JLIS.it

JLIS.it is a growing journal

Editorial board

JLIS.it, Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione, rivista dell'Università di Firenze, Dipartimento SAGAS, nasce nel giugno del 2010 come rivista open access, ospitata dal Cineca; nel 2015 assorbe il comitato scientifico e la direzione della rivista "Archivi & Computer"; diventa quadrimestrale e modifica il titolo in *Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione*; dal 2018 essa viene ospitata da EUM, Edizioni Università di Macerata, con il patrocinio finanziario del Master in formazione, gestione e conservazione degli archivi digitali di quella università.

JLIS.it ha colto l'occasione, che si è presentata circa un anno fa, di stabilire un'alleanza culturale originale, di grande rilievo, tra EUM e FUP, Firenze University Press: obiettivo è sperimentare modelli culturali e tecnologici innovativi e sostenibili, caratteristica distintiva della rivista. La collaborazione tra le due University Press aiuterà JLIS.it a essere maggiormente presente nel circuito della comunicazione scientifica internazionale come open peer-reviewed journal. Le azioni comuni e gli obiettivi da raggiungere sono ampi e da adeguare a un universo bibliografico e a un panorama editoriale in continua evoluzione. JLIS.it, infatti, corrobora il suo desiderio di essere sempre più vicina ai suoi lettori e di concepire nuove forme di riconoscimento e visibilità del lavoro svolto da autori e revisori.

JLIS.it inaugura una terza fase della sua storia: rinnova il board editoriale per rispondere meglio alle sfide impegnative di una rivista scientifica. Dal n. 2, 2022, il direttore fondatore sarà affiancato da un altro direttore che rinforzerà lo spazio per l'archivistica e alla *Lectio magistralis* si abbinerà il *Seminario JLIS.it* dedicato ai temi cari alla rivista.

JLIS.it, Italian Journal of Library and Information Science of the University of Florence, SAGAS Department, was born in June 2010 as an open access scientific journal, hosted by Cineca. In 2015 it absorbed the scientific committee and the board of "Archivi & Computer". It became quarterly and changed the title to *Italian Journal of Library and Information Science*. Since 2018 it has been hosted by EUM, Edizioni University of Macerata, with the financial support of the post-graduate Master on creation, management, and preservation of digital archives of that university.

JLIS.it took the opportunity, arousing about a year ago, to establish an original cultural alliance, of great importance, between EUM and FUP, Firenze University Press. The goal is to experiment with adopting innovative and sustainable cultural and technological models, distinctive features

---

JLIS.it

of the journal. The collaboration between the two University Press publishers will help JLIS.it to be more visible in the international scientific communication circuit as an open peer-reviewed journal. The actions and the objectives to be achieved are broad, adapted to a constantly evolving bibliographic universe and publishing landscape. JLIS.it corroborates its mission to be ever closer to its readers and to conceive new forms of acknowledgment and visibility of the work done by authors and reviewers.

JLIS.it inaugurates the third phase of its history: it renews the editorial board to better respond to the demanding challenges of a scientific journal. From no. 2, 2022 on, the founding director will be joined by another director who will reinforce the space for archives and the *Lectio magistralis* will be combined with the *JLIS.it Seminars*, dedicated to the hot topics for the journal.

JLIS.it

Universal Bibliographic Control today: preliminary remarks

Mathilde Koskas(a)

a) Bibliothèque nationale de France

ABSTRACT

Universal Bibliographic Control was formulated in the 1960s and 1970s, and was at the core of international bibliographic productions and exchange in the subsequent decades. However, in a digital ecosystem that is very different from the context in which it was born and thrived, it is important to examine what Universal Bibliographic Control means to the international bibliographic community, that is, the producers and managers of bibliographic – and authority – metadata, today. This paper is meant to invite discussion and reflections and to resonate with the various papers from the International conference on Bibliographic control in the digital ecosystem, organised by the University of Florence in February 2021. It focuses on the future of interoperability and the role of UBC in a democratic society, in the context of mass digital information, and its companion technologies.

KEYWORDS

Universal Bibliographic Control; Metadata; Interoperability; AI.

JLIS.it

The International conference on Bibliographic control in the digital ecosystem, organised by the University of Florence in February 2021, was a rare opportunity to examine in depth the idea of Universal Bibliographic Control (UBC), its relevance, the challenges it faces, in an information ecosystem that is so very different from what it was when the concept of UBC was first formulated and formalised in the 1960s and 1970s (Illien and Bourdon 2014, Guerrini 2021).

The scope and magnitude of the conference was of the kind that is seen maybe once a decade, and the last time the topic of Universal Bibliographic Control was examined by an international group of specialists and practitioners of comparable status was, to the best of our knowledge, at the joint open session of the Cataloguing, Bibliography and Classification & Indexing Sections and UNIMARC Strategic Programme of IFLA, the International Federation of Library Associations and Institutions, in Lyon, France, in 2014[1].

In 2021, of course, the topic, scope and international make-up of the conference was made all the more timely and relevant by the pandemic and its subsequent cancellation of international meetings. The international bibliographic community had been unable to meet in person to hold its usual discussions at the World Library and Information Congress (WLIC), IFLA's yearly international conference, in 2020. Meanwhile, the information landscape continued its fast-paced evolution, made, if possible, even faster by the increased importance of online communications during the pandemic.

The organisers built a very strong programme in the form of a dialogue between the Italian and international experiences, not confining it to the library world, either. Over the five half-days of the conference, going from the Italian point of view to the international and back gave participants a sounding board and a common thread in the dialectics of global and local, which was conducive to productive discussions.

This article is a formalised version of the opening remarks we were invited to give as Chair of IFLA's Bibliography section. It will examine what Universal Bibliographic Control means to the international bibliographic community, that is, the producers and managers of bibliographic – and authority – metadata, today. Like the speech it derives from, it is meant to invite discussion and reflections and to resonate with the various papers from the conference.

What is Universal Bibliographic Control to us?

During the aforementioned session on Universal Bibliographic Control in the Digital Age: Golden Opportunity or Paradise Lost? in 2014, the question was asked, "Did the digital tide knock UBC out?". Authors Françoise Bourdon and Gildas Illien noted the widely different ecosystem and the disparition of a formal governing body. But they also concluded that news opportunities had emerged which could form "the nodal point from which UBC's ideals may be invented once again". So, while Universal Bibliographic Control is admittedly almost 50 years old, has seen a deep evolution since its principles were first formally written down (Anderson 1974), and is now

---

[1] World Library and Information Congress: 80th IFLA General Conference and Assembly 16-22 August, Lyon, France. Session 86 - *Universal Bibliographic Control in the Digital Age: Golden Opportunity or Paradise Lost?* - Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Programme. http://library.ifla.org/view/conferences/2014/2014-08-18/315.html

JLIS.it

without a formal governing scheme, we think it is fair to say it still is the framework for our activity. In a way, the principles that presided over its conception and development have become such a fundamental part of bibliographic activities in libraries as to become perhaps largely implicit to librarians. Still, the activity of national bibliographic services contributes to Universal Bibliographic Control. Even if it changed, maybe beyond recognition, it hasn't disappeared but adapted. It shifted as our understanding of the underlying principles changed (be it the basing it on nations, the question of language, the objects of Universal Bibliographic Control themselves, not confined to just books anymore, or the focus on metadata rather than records). But we have noticed, in discussions with colleagues at home and internationally, that it still is the frame of reference for the production and distribution of bibliographic data, even when not explicitly invoked. The proceedings of the International conference on Bibliographic control in the digital ecosystem certainly confirm that observation.

If we examine some of the ways in which Universal Bibliographic Control and the ecosystem in which it exists have evolved, a few questions immediately arise, amongst which we will focus on the future of interoperability and the role of UBC in a democratic society, in the context of mass digital information, and its companion technologies.

Interoperability

The first question that comes to mind is that of interoperability. One of the founding principles of Universal Bibliographic Control is the sharing of bibliographic data. To that purpose, the tools of bibliographic exchange: standards (ISBD) and formats (MARC) were developed. Today's interoperability derives in part from these, but adapted to a completely renewed ecosystem of data exchange, relying on the internet and reaching far outside of the library world. An important change from the initial concept of Universal Bibliographic Control was the recognition of local needs, especially the need to access bibliographic information in one's own language. It modified the original concept, which was more concentrated. This is not just about the question of language, but in a broader sense, the taking into account of specific information needs and local cataloguing practices. Today, the international cataloguing code Resource Description and Access (RDA), which, interestingly, was not created under the auspices of IFLA, but gradually evolved into its current international scope and is now widely accepted as a major instrument for achieving the integration of bibliographic metadata in the semantic web, provides for the local, giving many options to cataloguing agencies on how to record and display information. Will this prove to be a problem on the global scale? Might these local rules become so fragmented as to constitute a challenge to interoperability? The reconciliation of local and global needs has been pointed out (Dunsire 2021[2]) as one of the main opportunities for the future of library metadata in the digital ecosystem. And indeed, if handled well, this challenge carries the seeds of opportunity. During the conference, one example of this came from the German-speaking countries' experience with the implementation of RDA, and the concept of a "Common core" (Behrens 2021).

---

[2] "The challenge for bibliographic control is the reconciliation of globalization and personalization via localization".

JLIS.it

Democratic role of UBC

Another important question revolves around knowledge and access to information, and their role in a democratic society[3].

Access to the entirety of the intellectual output of a society is an important condition of the democratic debate and a citizen's informed decision-making. This access is, of course, not possible if said output is not described with the appropriate metadata. Universal Bibliographic Control carries the promise to register, organise, and, ultimately, give access to everything. And while the promise is of course never entirely fulfilled, this objective has kept its relevance. Universal Bibliographic Control and the mass information era may have been said to be incompatible, but mass information (and its too painfully obvious pitfalls) underscores the need for the compilation and organisation of information that UBC strives for. That we are in an age of mass information doesn't mean that this work, this ideal, of Universal Bibliographic Control is useless, because it's hopeless, it means that, properly done, it is needed as much as it ever was, as long as we make it fit the new context. What librarians have to bring to the table is decades of reflection and practical experience of this encyclopaedic, universal idea (or ideal), and a framework and practices that have been in place for more than half a century. Whether we call it Universal Bibliographic Control or something else, the underlying principles of bibliographic information produced in accordance with international standards, in a way that is interoperable, accessible, and so on, are still there. We in the library world need to be careful not to let them cut us off from the world outside of libraries, but keep them more open than they have been in the past. With relevant and continuing adaptations, Universal Bibliographic Control remains a useful framework in today's digital ecosystem.

Shifting tides

We are shifting from distributed bibliographic control to shared entity management. This conceptual evolution comes with a reevaluation of libraries' scope of action. In the moving from bibliographic and authority records to entities, librarians have to ask themselves which entities libraries should take responsibility for, what level of quality is promised to users for each entity, and, crucially, how to work with other metadata producers, especially for what libraries can't take complete responsibility for (Leresche 2021, Boulet 2021).

In this new ecosystem, another protagonist has appeared: the machine, in the form of artificial intelligence, whose possibilities libraries and the metadata world is only starting to explore. Experiments around the world, such as Annif and Finto AI (Mödden 2021, Suominen 2021), show both the great potential of these technologies and the great investments (of skill, time, energy and money) they require. Ethical questions will also have to be addressed. Like all technological advances (for example the computerisation of libraries), it will turn out to be useful in its place, not so much reducing the human workload as shifting it. We learned from previous instances that it's important not to embark on technological choices that are specific to libraries, cutting our metadata off from the wider world. This is a pas de trois, involving libraries, the wider metadata and information communities, and the machine, not a pas de deux.

---

[3] Schreur 2021; Guatelli 2021; Bourke 2021.

JLIS.it

Most of us feel that we are living in very chaotic times, professionally speaking. At the French National Library, for instance, we are working at the same time on a new cataloguing code (RDA-FR, a French version of RDA), its application profiles, a new format[4], and a new cataloguing application, to say nothing of training, etc. This is actually a global issue, as this is happening all over our institutions right now, France being no exception. It is quite challenging, but also potentially very fruitful. As the various projects' progress is parallel in terms of temporality, each one informs the others, in a dialogue, in terms of method. Chaotic it may feel, but from chaos springs creation, as the initiatives and experiments presented at this conference abundantly proved.

---

[4] INTERMARC Next Generation, see Peyrard and Roche 2018.

JLIS.it

References

Anderson, Dorothy. 1974. *Universal Bibliographic Control: a Long Term Policy, a Plan for Action*. Pullach bei München: Verlag Dokumentation.

Behrens, Renate. 2021. "Standards in a new bibliographic world – community needs versus internationalisation". Paper presented at the *International conference on Bibliographic control in the digital ecosystem*, Florence, Italy, February 08-12, 2021.

Boulet, Vincent. 2021. "Towards an identifiers' policy: the use case of the Bibliothèque nationale de France". Paper presented at the *International conference on Bibliographic control in the digital ecosystem*, Florence, Italy, February 08-12, 2021.

Dunsire, Gordon and Mirna Willer. 2014. "The local in the global: universal bibliographic control from the bottom up". Paper presented at: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 86 - Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Programme. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. Accessed April 14, 2021. http://library.ifla.org/id/eprint/817.

Guatelli, Fulvio. 2021. "Maximising dissemination and impact of books: the scientific cloud". Paper presented at the *International conference on Bibliographic control in the digital ecosystem*, Florence, Italy, February 08-12, 2021.

Guerrini, Mauro. 2021. "Opening remarks". Paper presented at the *International conference on Bibliographic control in the digital ecosystem*, Florence, Italy, February 08-12, 2021.

IFLA Professional Statement on Universal Bibliographic Control, December 2012. Accessed April 14, 2021. https://www.ifla.org/files/assets/bibliography/Documents/ifla-professional-statement-on-ubc-en.pdf.

Illien, Gildas and Françoise Bourdon. 2014. "UBC reloaded: remembrance of things past, back to the future". Paper presented at: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 86 - Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Programme. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. Accessed April 14, 2021. http://library.ifla.org/id/eprint/956.

Leresche, Françoise. 2021. "Rethinking bibliographic control in the light of IFLA LRM entities: the ongoing process at the National library of France". Paper presented at the *International conference on Bibliographic control in the digital ecosystem*, Florence, Italy, February 08-12, 2021.

Mödden, Elisabeth. 2021. "Artificial intelligence, machine learning and DDC Short Numbers". Paper presented at the *International conference on Bibliographic control in the digital ecosystem*, Florence, Italy, February 08-12, 2021.

Peyrard, Sébastien and Mélanie Roche. 2018. "Still Waiting for That Funeral: the Challenges and Promises of a Next-Gen INTERMARC". Paper presented at: IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies in Session 141 - Cataloguing. Accessed April 14, 2021. http://library.ifla.org/id/eprint/2204.

Schreur, Philip. 2021. "*I'm as good as you*": the death of expertise and entity management in the

JLIS.it

age of the Internet". Paper presented at the *International conference on Bibliographic control in the digital ecosystem*, Florence, Italy, February 08-12, 2021.

Suominen, Osma. 2021. "Annif and Finto AI: developing and implementing automated subject indexing". Paper presented at the International conference on Bibliographic control in the digital ecosystem, Florence, Italy, February 08-12, 2021.

World Library and Information Congress: 80th IFLA General Conference and Assembly, 16-22 August, Lyon, France. Session 86 - *Universal Bibliographic Control in the Digital Age: Golden Opportunity or Paradise Lost?* - Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Programme. Accessed April 14, 2021. http://library.ifla.org/view/conferences/2014/2014-08-18/315.html.

JLIS.it

Conference BC 2021

Josep Torn[a]

a) European University Institute

ABSTRACT

This article is a compendium of some of the presentations made during the BC2021 conference.

KEYWORDS

Bibliographic control; Metadata; Research data; Open access; Authority control; Research libraries; National libraries; Musicology collections; Artificial intelligence.

JLIS.it

The Universal Bibliographic Control (UBC), as indicated by professor Mauro Guerrini (Università degli studi di Firenze, IFLA Bibliography Section, chair of the Conference) in his opening remark, is an exercise that intellectually begins, at least, with Conrad Gesner's Bibliotheca Universalis. We are confronted with a panorama that allows more than ever to advance towards IFLA's objective of making catalogue records available immediately, an exercise in which libraries have always excelled in its two aspects: thoroughness in document description and willingness to share knowledge in any of its stages.

The Conference on Universal Bibliographic Control (BC 2021) touched on key aspects for, in a digital ecosystem, making the maximum of resources available, opening interesting debates on new standards (or the evolution of current ones). Some aspects, formats or objects take on greater significance such as data, authority control, multilingual collections, or artificial intelligence. These aspects, although key, are not new to the librarians. As Mauro Guerrini reminds us, already in 2014 during the IFLA conference in Lyon (France) he raised the key question that librarians have still do not solved: "Digital age: Golden opportunity or Paradise lost?"

Mathilde Koskas (Bibliothèque nationale de France, IFLA Bibliography Section, chair) proposed the ideal departing point, starting from the relationship between local work (Italy, for the case) as the basis for a, step by step, more global approach. Koskas raised key questions for the UBC, such as the role of the national libraries in this ambition. Both, Guerrini and Koskas, emphasised basic aspects to UBC today such as interoperability, multilingualism or the international cataloguing practices in local. The democratic role of UBC overcomes the barriers that mass information seems to want to impose as more universal, since it compiles information in a complex context where *mass* means quantity without quality assessment or veracity control. Mathilde Koskas proposes a [maybe] new role of responsibility for the librarian – role that she opposes precisely to that of the automated systems, where we still have to learn what kind of results they give or will give and what benefit they offer in terms of knowledge organisation.

Renate Behrens (Deutsche Nationalbibliothek, Germany) opened fire with a central issue: standards. Standards and their meaning in this new bibliographic framework. Behrens described the current library environments as challenging, because of the need for (still another) transition, as well as promising because of the role of librarians as mediators that guarantee participation in social development. Standards help on this objective by "putting the world of things in order", but as Behrens indicates "standards do not establish the order of the things themselves". Standards are crucial for those libraries that want to exchange information and share content, or for those that have a common goal that they want to advance on. Behrens reminds us of the importance of keeping the standards up to date, otherwise, they lose the aim for what they were created (and maybe even all the work behind them).

Standardisation was also the key aspect that Andrew MacEwan (British Library, UK) touched upon. He focused more on authority control and name identifiers. His presentation, about the International Standard Name Identifier, started by posing for discussion the huge amount of metadata models that libraries use today that, for sure, make life easier to many but that present a complex playground for the interconnection of knowledge. MacEwan did not see a big challenge though, due to the variety of metadata silos from where crosswalks are created, but he raised concerns about the quality of the metadata and the need to count on *this quality* at the beginning of the supply chain.

JLIS.it

The International Standard Name Identifier (ISNI) is an ISO standard that the British Library uses as a registration agency as a tool to unequivocally identify creators that can play different roles along with their creative career. But despite the fact that ISNI presents itself as a standard and it is precisely an ISO, Andrew MacEwan warned of important challenges in order to go further, such as to become a tool for the collaboration with LoC or to be adopted by all UK publishers.

The British Library was not the only national library to address the topic of music in relation to bibliographic control. The Bayerische Staatsbibliothek (Germany) has one of the largest collections of music and musicology in the world, and Klaus Kempf explained how the application of the RDA in different specific cases has been implemented in the Bavaria National Library. The Bayerische Staatsbibliothek has brought the search of music documents to another level, with its project Melody Search; an optical music recognition search engine.

The Biblitoteca Nazionale Centrale di Roma (Italy), for its part, added solutions to working with digital materials, including also Open Access objects. The paper, by Fabio D'Orsogna and Giulio Palanga, was centred on the example of a final front end that uses form standards for description, but also the long path still pending to walk in collaboration with other libraries. It was a constant by the different national libraries that presented at BC2021 to do not only describe their internal procedures or methods, but to illustrate the results by using clear front-ends where professionals and users see the application of standards or models; which is more than welcomed.

Continuing with national libraries, Osma Souminen presented an example on how to bring bibliographic description to another level, combining artificial intelligence (AI) with manual text code used in classification. The National Library of Finland has created an Open Source solution, *Annif*, that has evolved into Finto AI. Finto AI integrates semi-automated subject indexing into metadata workflows, a tool that it is already used by libraries in Finland. Introducing automated subject or bibliographic description is not the sole objective of the Finnish. Also in Germany, Elisabeth Mödden (Deutsche Nationalbibliothek) and her team have worked on the automated assignment of Dewey Decimal Classification numbers.

For Renate Behrens, Deutsche Nationalbibliothek, collaboration continues to be crucial. National libraries do not only have the role in guiding libraries and librarians of their nations, but the commitment to seek collaborative solutions in relation to the use of standards, in that case those used in bibliographic description.

Collaboration – when applied to national libraries – means, precisely, internationalisation. Vincent Boulet, (Bibliothèque nationale de France) mentioned the need to define identifiers' policies, be them for international – again – or even local models. And still from the BNF, Françoise Leresche recalled the transition from ISBD and Unimarc to new models like LRM that IFLA has sponsored. The BNF is a provider of metadata for cataloguers beyond the walls of the national library and beyond the boundaries of France.

We also saw how national libraries are concerned about final services, for which they rely on bibliographic control to assure the quality of the information involved in services. Oddrun Pauline Ohren (Nasjonalbiblioteket – National Library of Norway) addressed the need for solid use of bibliographic control standards to be able to cover "every corner of Norway" with digital material, media podcasts or streaming events (among others), straddling – thus – the back office and the front office of library services.

Professionals from academic libraries addressed as many different issues as the national libraries'

JLIS.it

ones. Tiziana Possemato (Università di Firenze) put together the Universal Bibliographic Control (UBC) with the semantic web. She advocates for a dialogue between systems in the form of the exchange of records that overcomes cultural, linguistic or geographical limits. Similarly, the University of Alberta Library, represented by Ian Bigelow and Abigail Sparling, presented the conversion of standards (RDA and MARC) to BIBFRAME as examples of collaborative innovations. There was also time for research datasets, not covered by any other speaker, Thomas Francis Bourke (European University Institute Library, Italy) explored how the bibliographic control function has been expanded to embrace research data in the social sciences and humanities. Bourke claims that data librarians need to work closer to research data management (RDM) units by using formal bibliographic control functions. The relation between wikidata and UBC was discussed by Lucia Sardo and Carlo Bianchini (Universities of Bologna and Pavia [Italy], respectively). Sardo and Bianchini offered a theoretical but also a practical approach, arguing that wikidata shows that we need to overcome the only approach of the national libraries to embrace more co-operative approaches.

Another crucial and interesting aspect addressed during this edition of the Conference on BC was multilingual collections and UBC by Pat Riva, from Concordia University Montréal (Canada). Institutions like Riva's, with users that represent a variety of native languages amongst their community, may find it difficult to search by using the library discovery tools in their own languages when the description of the objects has been solely made in one of the languages of the society in play (the predominant one). CUM has integrated some strategies by using linkages between authority files in English and French.

We have had red flags raised about the wrong or too limited use of metadata that librarians do. Richard Wallis warned us that, while many other actors in the information industry use metadata to make others aware of their resources, libraries tend to hide these metadata in the back-office. With this practice, we lose potential users and customers.

And as a final remark, and leaving some other interesting presentations unmentioned, as Michele Casalini (Casalini Libri) said talking about the future for an international audience, there is the need for connected services and automatic processes to help enrich the information we provide to our users. This challenge needs to be addressed not only with interoperability but with international cooperation.

JLIS.it

Universal bibliographic control in the digital ecosystem: opportunities and challenges

Mauro Guerrini[a]

a) Università degli Studi di Firenze

ABSTRACT

The idea of universal bibliographic control (UBC) has been of interest for centuries in the history of cataloguing and is based on the humanistic ideal of sharing recorded knowledge produced anywhere in the world. In the contemporary era, IFLA has played a central role, stimulating national bibliographic agencies and other institutions to promote standards and collaborations that go beyond the national sphere, leading to multicenter and even more cooperative bibliographic control. The tradition of cataloguing also grows and is enriched by the dialogue with different communities and users' groups. The free reuse of data can take place in contexts very different from the original ones, multiplying for all the opportunities for universal access and the production of new knowledge: the UBC, therefore, looks at interoperability and flexibility in the dialogue with the various communities of stakeholders and with the cultural institutions.

KEYWORDS

Universal Bibliographic Control; UBC; Cataloging.

JLIS.it

Culture is the only asset of humanity that, when divided between us all, becomes greater rather than smaller.
Hans-Georg Gadamer

As a "non-commercial public space" (IFLA Global Vision) – not only in a literal sense – libraries play a fundamental role also in the digital ecosystem
Conference BC2021

Bibliographic control: a central topic in LIS

The idea of universal bibliographic control has been of interest for centuries in the history of cataloguing, and it is based on the humanistic ideal of sharing collective knowledge in every part of the world. It probably began with Conrad Gesner's *Bibliotheca Universalis* (1545–1549), the catalog of all printed books published up to that time in Latin, Greek, and Hebrew. Gesner called 'Universalis' his work, pursuing the goal of maximum bibliographic coverage in relation to the concrete literary reality of his time. His universal bibliography included a catalog for authors' names, and a catalog for general as well as specific subjects (*loci*). Gesner established the connotations of the scientific and literary heritage and established the characteristics of indexing logic using four categorical levels: author, work, text, and edition.[1]

In the contemporary era, IFLA has played a central role in the realm of Universal Bibliographic Control (UBC) by bringing together national bibliographic agencies and other institutions to promote standards and collaborations in this area. This also includes the work of promoting conferences and publishing texts and documents.[2] From 1990 through the 1st of March 2003, the Deutsche Bibliothek hosted the IFLA Universal Bibliographic Control and International MARC Core Activity (UBCIM),[3] demonstrating the direct connection between UBC and technologies. For years IFLA has edited "IFLA Series on Bibliographic Control". In particular, one book in that series entitled "National Bibliographies in the Digital Age: Guidance and New Directions", edited by Maja Žumer in 2009,[4] continues to be a fundamental reference text. A statement reaffirming IFLA's commitment to UBC was endorsed by the Professional Committee in December 2012. Initiated by the Bibliography Section, that statement was also supported by the Cataloguing Section and the Classification and Indexing Section.[5] The WLIC of Lyon in 2014, included in the programme a seminar entitled "Universal Bibliographic Control in the Digital Age: Golden Opportunity or Paradise Lost?"[6] It was planned by the Cataloguing Section, with the Bibliography Section, the Classification Section, and the UNIMARC Strategic Programme.

---

[1] (Sabba 2012).

[2] (Anderson 1974); (Davinson 1975).

[3] https://archive.ifla.org/ubcim/.

[4] (IFLA 2009).

[5] https://www.ifla.org/publications/node/7468.

[6] Monday, 18 August 2014; see Session 86, http://library.ifla.org/id/eprint/817/.

JLIS.it

Also, back in 2001, the Library of Congress organized the "Conference on Bibliographic Control for the New Millennium",[7] celebrating a significant anniversary precisely with this theme. The Library of Congress established an independent Working Group on the Future of Bibliographic Control that published the report entitled "On the record" in 2008.[8]

As we can see from these recent events, bibliographic control is central to the history of cataloguing and to the history of libraries themselves.

The concept of Bibliographic Control has changed and still changing radically, because the bibliographic universe and technologies are radically changed; and resources, actors, standards, and practices will presumably change further. It necessary, therefore, to explore the new boundaries of bibliographic control, in fact, the digital ecosystem.

Text and metadata as paradigm of bibliographic control

For centuries, a text (whether manuscript or printed) was identified by the physical volume. Today, 'work' is at the center, and increasingly its content can be presented and enjoyed in many forms. For example, a reader can choose between paper and e-books, based on his or her reading preferences. This content is now usually accompanied by a set of metadata. Metadata has become the protagonist of communication on the web; metadata is today the paradigm of bibliographic control. Some of the consequences are already evident. For example, the quality metadata of a resource contribute to its knowledge, enhancement, and success.[9]

The process of metadata creation for bibliographic resources starts with the creators of those resources – obviously providing the content –, and, in the modern era, usually providing the title, and some basic metadata; then, the publishers add their metadata, including some standard identifiers, an important step in the bibliographic control in the digital ecosystem. The process of metadata creation continues through the intellectual contribution of the cataloguers of the bibliographic agencies. Considerable is the initial investment in the creation of metadata based on authoritative sources.[10]

From the model of universal bibliographic control based on the centrality and exclusivity of the national bibliographic agencies, we are moving on to dynamic and shared bibliographic control. In the digital world, this is configured as a process of data reuse and enrichment, linking single data elements. In an evolving ecosystem, the international dimension is the virtual space where stakeholders meet. In this context, libraries, and in particular, the national libraries, no longer have the monopoly of bibliographic control. This poses an intellectual and operational challenge to library institutions. However, libraries, library networks and bibliographic agencies still play an important role, in particular, through strong collaborations among themselves, through their role as true protagonists of the standards of bibliographic control, standards flexible and at the same time binding and reliable. Still, libraries remain an essential part of the digital ecosystem.

---

[7] Library of Congress, "Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium" https://www.loc.gov/catdir/bibcontrol/.

[8] https://www.loc.gov/bibliographic-future/.

[9] (Guatelli 2020).

[10] As an added aspect, metadata can serve as an antidote to even fake news; cfr. (Bredemeier 2019, 384 and so on).

JLIS.it

What are the consequences of digital transformation for library catalogues, and work processes in metadata creation? What is the function of repositioned and reconfigured catalogues on the web? Understanding how texts are conveyed today requires cultural awareness and professional training: this is the basis of the process of literary and conceptual analyzing the resource. These two aspects – awareness and training – should be common to the training of other actors involved in the process, who serve as mediators of the knowledge process.

Beyond tradition

The data models and the semantic web paradigm invite us to go beyond that aspect of the cataloging tradition that entrusted only the bibliographic agencies with the role of authoritative producers of quality registration. Data models and the semantic web paradigm invite us to go beyond the cataloging tradition. That tradition provided for homogeneous descriptions for all the libraries. The contemporary perspective foresees the participation of new and different actors. In addition to libraries and librarians, other institutions (publishers, distributors, private agencies, universities), and professionals (archivists, museum professionals) are contributing to the recording and enrichment of metadata and authority files. In those context, libraries still play the role of intermediary with the other major producers of metadata. The participation of several actors is very positive, and everyone is invited to find a new balance between their different methodological and cultural traditions to pursue a common goal: the cooperative editing of quality metadata, possibly in open access. The best cataloging tradition in the completely new collaborative context is therefore maintained and indeed enhanced.

Another consequence is that the relationship between libraries, publishers and distributors becomes more strategic, because the publishers are the first, after the creators themselves (in the modern era), who should create the metadata of a resource, and later, that metadata is enhanced by libraries for the part that concerns libraries. Libraries feel, with particular responsibility, the issue of the shared construction of quality data, by virtue of the principles of precision, accuracy, and social sharing of the cultural heritage that have characterized their history.

Bibliographic control today is, therefore, multicentric, and even more cooperative than in the past. National bibliographic agencies maintain and reinforce their role in quality control of metadata and authority control, through the maintenance of fundamental tools, such as VIAF (Virtual International Authority File) and through support of international identifiers such as ISBN (International Standard Book Number), ISSN (International Standard Serial Number) and ISNI (International Standard Name Identifier), that are part of broader international cooperation and authority control projects.

VIAF and ISNI are different projects: VIAF is an international collaboration that supports a shared authority file; ISNI is a name identifier and a system for recording those numbers that define it. VIAF, in particular, provides authoritative services that reliably identify agents, places etc., and the works associated with them in the global registered knowledge network. Its philosophy is inspired by promoting all cultural perspectives equally, including all languages and scripts, and simplifying the work of bibliographic agencies and libraries. Many libraries and bibliographic agencies collaborate in sustaining these authoritative resources for the benefit of users everywhere.

JLIS.it

The greater the accuracy of the data, the greater the benefits of using those authoritative sources. By aggregating and linking data, these sources for authority control can bring greater interoperability to the galleries, library, archival, and museum community (GLAM) as well as the publishing and book dealership industries.

The form of the name: conditioned by the cultural and linguistic context

The choice of form of a name associated to an entity is always culturally founded, but the selection of the preferred form of a name is, in many cases, complex, and depends upon the cultural and linguistic context in which that name is used. In the past, the bibliographic traditions of the Western world were privileged, but now the global dimension of communication changes all parameters. In the global cultural environment (as opposed to a single library's catalogue), there has been the important acknowledgement that there is no single form of an author's name that must be used by everyone. The choice of the form of a name to be displayed is conditioned by the cultural and linguistic context within which the dataset for that name is placed. IFLA LRM recalls that a named entity can have different *nomen*, all valid (e.g., Léonard de Vinci in France and Leonardo da Vinci in Italy; Cicero in a specializing library in Latin literature and Cicerone in a public library). The goal is to overcome the geography and dominance of a cultural area, and to respect the cultural and linguistic traditions of each Country, and of each individual cultural community in the solutions adopted.

The mechanism of "reconciliation" of the different forms with which an entity is known and identified in a global context (for example, the creator of a work), brought together in a group of variant forms, all recognized, becomes the principle for new ways of sharing information. The entity reconciliation process produces a cluster: it is a grouping of the different variant forms referable to the same entity; this entity is known in various nomen in different cultural, linguistic, geographical, domain contexts; all valid, usable and actually used variants. Linking various identifiers is of strategic importance. In all entity identification projects that make use of the reconciliation (or clustering) mechanism, it is customary to assign an identification to the recognized entity; identifier that connects to other identifiers assigned to the same entity in different contexts, and all valid. The clustering mechanism starts from the assumption that all forms of a name used in the global context have equal dignity; there is no particular preference for one or the other form. The context of belonging (the source from which that variant form of the name comes) and the need for use (the target that recalls that name) define each time the choice of the form to be considered the preferred "conditioned" form of the name. This is motivated by the desire to enrich the dataset, and to offer the reader as many channels as possible to reach the goal; this is the pragmatic and functional purpose of being able to identify, select and obtain the resource. The identifiers allow both the explication of the equivalence function of the forms of the cluster and the connection of the cluster to other clusters relating to the same entity. The choice of the preferred form of the name, the structuring of the string (according to syntactic rules known in the past only to catalogers), lose importance in the face of the practical need to create multiple and equivalent retrieval channels for the same resource.In the context of Universal Bibliographic Control, there remains the need to offer a form as a result of a national or cultural or linguistic choice; this is also achieved

through information presentation mechanisms linked to the cluster: the data on the "provenance" of the information (given on the source that produced the information) can be used in a double meaning:

- the more traditional: source that generated the information and that defines, within a cluster, which form is to be presented as preferred in a given context;
- that of the applicant target (Provenance of the applicant) which, on the basis of its own specific research need, guides the selection of the preferred form (also in this case, therefore, preferred in the specific context of the research).

Therefore, the cluster of variant forms is fundamental passage from Bibliographic Control intended as control of strings and access points to the more complex concept of entity identification, through different and variant identities with which it can be expressed. The choice of the form of the name and the linking of variants in clusters enhances the concept of universal bibliographic control that respects cultural variations for the display of names.

The tradition of cataloguing grows and enriches in dialogue with different communities and groups of users. The free reuse of data can take place in very different contexts from the original ones, multiplying for all the opportunities for universal access and for the production of new knowledge. The concept of cultural heritage values is a living idea.

The great changes brought on by the use of metadata have led to new perspectives on bibliographic control. UBC now contemplates interoperability and flexibility in dialogue with the various communities and with institutions of registered memory.

Who knows what the future will bring us? Perhaps, we are still at the beginning of the digital revolution. Precisely in the field of metadata and authority control, we could expect developments and surprises from alternative technologies on machine learning or artificial intelligence, a tool that promises to be very useful; a tool that takes nothing away from the cataloguer's judgment, which remains a fundamental intellectual activity.

JLIS.it

References

Anderson, Dorothy. 1974. *Universal Bibliographic Control: A long term policy, a plan for action.* Pullach/München: Verlag Dokumentation.

Bredemeier, Willi. 2019. *Zukunft der Informationswissenschaften: Hat die Informationswissenschaft eine Zukunft? Grundlagen und Perspektiven. Angebote in der Lehre. An den Fronten der Informationswissenschaft.* Simon Verlag für Bibliothekswissen.

Davinson, Donald Edward. 1975. *Bibliographic Control.* London: C. Bingley.

Guatelli, Fulvio. 2020. «FUP Scientific Cloud e l'editoria fatta dagli studiosi». *Società e storia* 167. doi:10.3280/SS2020167008.

https://archive.ifla.org/ubcim/.

http://library.ifla.org/id/eprint/817/.

https://www.ifla.org/publications/node/7468.

https://www.loc.gov/bibliographic-future/.

IFLA. 2009. *National Bibliographies in the Digital Age: Guidance and New Directions.* A cura di Maja Žumer. IFLA Series on Bibliographic Control 39. München: K.G. Saur.

Library of Congress, "Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium" <https://www.loc.gov/catdir/bibcontrol/>.

Sabba, Fiammetta. 2012. *La 'Bibliotheca Universalis' di Conrad Gesner, monumento della cultura europea.* Roma: Bulzoni.

JLIS.it

Standards in a new bibliographic world

Renate Behrens[(a)]

a) German National Library

ABSTRACT

Jointly developed and agreed standards are essential for description and exchange of data on cultural assets. We are at a turning point here. Standards with broad acceptance must move away from strict sets of rules and towards framework models. To meet this challenge, we need to fundamentally rethink the conception of standards.

Cultural institutions hold treasures and want to make them accessible to a wide range of interested parties. What was only possible on site not so long ago, now also takes place in virtual space and users worldwide can access the content. To make this possible, all resources must be provided with sufficient and sustainable metadata. Many sets of rules and standards can do this and aim to make the exchange of data as international and large-scale as possible.

But does this also apply to special materials? Is a lock of hair to be recorded in the same way as a book, or is an opera to be redorded in the same way as a globe? By now, it is clear to everyone involved that this is not the case. Far too much expertise is required for this, which is not available in the breadth of cataloguing. This is quite different in the special communities, where this expertise is available and many projects and working groups are working intensively on the relevant topics. In order to bundle these approaches and enable more effective cooperation, the colleagues must be networked and embedded in a suitable organisational structure. This is the only way to achieve results that are accepted by a broad range of users and at the same time are sustainable and reliable.

This article is intended as an introduction to a future discussion and does not aim to provide answers.

KEYWORDS

Standards; Internationalization; Cataloguing; Special Resources; Objects; Collections; Cooperative Cataloguing; Data Exchange.

JLIS.it

What is the new bibliographic world?

The world of information and documentation institutions has changed dramatically in the past few years. Information must be made available both quickly and reliably. The speed at which information flows has increased exponentially, whereas the durability of the data has decreased considerably. The worlds of academia and research produce and distribute information in vast quantities and update it virtually in real time. New methods of production and distribution are capable of making this information available, reusing it, changing it and reintroducing it to the data cycle, all within a very short period of time.

The basis for this was and remains the technical innovations of recent decades, which made these processes feasible in the first place and whose effects have been so profound that they have also transformed society. In those areas of the world where democracy is established, all levels of society – regardless of sex, age or world-view – gained access to knowledge and information. Life-long learning and education became available to far more people and are now taken for granted by the younger generation.

At the same time, this achievement also necessitates more stringent quality control. Data can be altered, falsified and reintroduced into the information cycle with the same speed that they can be produced in the first place. So-called fake news has become an ignoble part of our global communication in recent years.

Every information and documentation institution must reinvent itself in this new environment. The traditional tools used in libraries, archives and museums are no longer sufficient to the task. These tools are no longer adequate for administering and controlling the global data streams with the desired quality or speed, and the large quantities of data can no longer be tackled with conventional means. It is essential to create synergies and intensify or establish international, interdisciplinary cooperation. To this end, it should be self-evident that the efficacy of the old, familiar tools and approaches must be re-examined.

What role can international standards play in this context?

Standards provide the foundation for the generation and functional exchange of data. Even communities that seem to be highly independent will sooner or later reach a point where they require shared agreements and regulations in order to ensure the interchangeability of data and maintain a certain level of quality. Effective and contemporary standards can accelerate the editing and generation of data and increase efficiency in the further use of data. To achieve this, however, these standards must be updated continuously and adapted to the current circumstances. General standards that are adapted by the respective user communities to their specific needs can be of benefit in this context, but also require a large degree of initiative on the part of the respective community. Modular standards are easier to work with and more flexible in their application. In many instances, a minimum degree of consensus is all that is required to ensure the exchange of data. Special requirements can be added in dedicated modules, which in turn are then further developed by experts in the respective field. In light of the aforementioned developments, rigid frameworks that contain fixed rules and are heavily text-based have proved to be no longer fit for purpose.

In this context, authority data have become particularly significant. They are a tried-and-tested

JLIS.it

tool in libraries and are labour-intensively administered there – within the Integrated Author-ity File (GND) in German-speaking countries, for example, or using the Library of Congress Authorities in Anglo-American countries – and, in some instances, collated within intraregional data such as the Virtual Authority File (VIAF). However, the importance of authority data has further increased as a result of increasingly interdisciplinary collaborations. Authority data e.g. for individuals and geographic entities are the smallest common denominators for the collaboration across different communities. Yet the altered circumstances have also resulted in fresh challenges. In addition to expanding the vocabulary, new concepts must be developed and a shared definition created for entities that have hitherto been imbued with different meanings and the subject of diverging interpretations. For example, the term "work" is interpreted differently in the world of archiving than it is in library-related contexts.

Who are the stakeholders in this new bibliographic world?

As described in the preceding section, data-administering cultural institutions are an essential part of our society. This is nothing new; for centuries now, libraries, archives and museums have been re-sponsible for the preservation and administration of our cultural heritage. Yet this task has long been regarded as an activity exclusively for the benefit of a select clientele. By contrast, modern cultural institutions regard themselves as habitats, sometimes to an extent that exceeds their legal mandate. New library and museum buildings around the world stand as testimony to this fact. Yet it is not just the external appearance of cultural institutions that has to adapt to these new circumstances, but also the products and services they provide. However, this adaptation must occur not only in line with the respective institution's own community, but also on an interdisciplinary basis.

Unlike 50 or 100 years ago, say, the updating and new development of standards in the sphere of information science requires the input of expertise from many different areas. Technical expertise is a given in this context; however, sociological and socially relevant aspects must also be fac-tored in. If standards are to continue adhering to the International Cataloguing Principles (ICP)[1], then users' search habits and the reliability of the generated data must be included amongst the key criteria. Democratic methods for developing standards are also desired today, which gener-ally increases the development period but also ensures considerably greater acceptance. Ideally, standards should already be considered from different perspectives in terms of their intended use, target audience and applicability before they are actually developed or updated. Especially when it comes to implementing theoretical concepts and models, attention must be paid to their practical relevance, and the expertise of colleagues working in user communities and educational institutions sought. Sensibly, global feedback phases are no longer a rarity, and an interdisciplinary perspective should become a matter of course.

Sound and practicable organisation is required in order to bring together these different play-ers. In general, libraries have the requisite standardisation committees at their disposal and have gained lots of relevant experience over the decades. Examples of such collaborations will be de-scribed in the next section.

---

[1] Cf. https://www.ifla.org/publications/node/11015.

JLIS.it

What role do the user communities play?

Due to changed circumstances, the user communities play a greater role in the development of standards than was previously the case. Flexible standards must be repeatedly analysed to ensure that they are up to date, and continuously amended. The assumption that the adoption of national or international standards could negate the need for any standardisation work of one's own has proved false. A comprehensive and international standard cannot meet the needs of the often very heterogeneous communities, but merely provide the basis for local and subject-specific adaptations. What is required is a group of experts in the areas of data generation, the further use of data by community members and technical parameters. This task is resource-intensive and expensive but can result in efficiency-savings when narrowing the broad scope of standards and their application. This is because the needs of, for example, those performing cataloguing work are known and can be taken into consideration when adapting the standards. In future, this task will require the establishment of a greater knowledge-base and expertise in the training of specialist staff.

Examples

We will now provide three examples to further illustrate the requirements outline above. These are standards that originate from very different traditions and areas of application, and yet feature certain commonalities.

Rules on Cataloguing Authority Data in Archives and Libraries (RNAB)[2]

This standard was first published in 1997 under the name "Rules on Cataloguing Autographs and Legacies" (RNA) and is used for these kinds of material by many archives and libraries. Since 2015, the standard has been painstakingly revised and was first published on the website of the German National Library in 2019. The organisation of this standard is regulated in a dedicated co-operation agreement between the Austrian National Library, the Swiss National Library, the Berlin State Library and the German National Library. The update was carried out by a thematic working group of the Committee for Library Standards[3] and underwent a comprehensive assessment procedure performed by colleagues working in archives and libraries.

In terms of its content, the standard has predominantly been optimised for use in literary archives. Alongside the actual revision of the rules, the circumstances of the institutions using the standard have also been taken into consideration at every stage. Thus the RNAB have deliberately been kept brief, dispensing with any complicated theoretical models. This was done in awareness of the fact that many institutions wishing to process this material do not have staff trained in Library Science at their disposal and that the cataloguing work has to be performed by other employees in addition to their primary tasks. For practical reasons, the standard was published at a time when it was clear that it would shortly require further revision due to changes in the fundamental model.

---

[2] Cf. https://www.dnb.de/EN/Professionell/Standardisierung/Standards/_content/rnab_akk.html.
[3] Cf. https://wiki.dnb.de/display/STAC/AG+RNAB.

JLIS.it

The feedback from the user communities has been uniformly positive and vindicates the practical approach of the RNAB.

3R Project for DACH Libraries

The international Standard Resource Description and Access (RDA)[4] was first introduced in German-speaking countries in 2014 for the cataloguing of authority data, and then for bibliographic data in 2015. Due to changes in the standard, a project for the necessary adaptations was set up in 2020. This so-called 3R Project for DACH Libraries implements the above-described community-centred approach to standards. By means of a cataloguing handbook as a web-based tool, the rules of the RDA are being prepared for the user communities in German-speaking countries and documented in a cataloguing handbook. This handbook will be composed of three sections: the descriptions of the elements, the descriptions based on resource types, and general instructions and assistance. As an end-product, it will provide the foundations for the practical cataloguing of data in the respective institutions, but also form the basis of staff training and induction. The provision of the handbook as a web tool opens up many options for subsequent use and for institutions to compile their own information and examples with links to the original RDA standard. The project is set to be completed by late 2022 and introduced within the institutions by training staff in the use of the revised standard. The DACH cataloguing handbook is being developed by the cataloguing expert group[5], a group of experts from library unions, public libraries and national and state libraries. The work has been commissioned and organised under the aegis of the Committee for Library Standards.[6] Specialist materials such as art books, graphic materials and audio-visual media have been incorporated into this process. The thematic working groups of the Committee for Library Standards are responsible for this task and will participate in the resource-description work from late 2021 onwards. The new cataloguing handbook will be documented in a web-based tool modelled on Wikibase. The work is being carried out within the DNB as part of an in-house documentation project.

International Standard Bibliographic Description (ISBD)[7]

Within the world of libraries, the ISBD is a very well-known and globally used standard issued by the International Federation of Library Associations and Institutions (IFLA).[8] It was first published in 1971 and has been revised and expanded many times since then. The current version is the Consolidated Edition from 2011.

The ISBD seeks to provide a basic standard for as many different applications as possible in different environments and regions. Based on this fundamental principle, the aim is to make the exchange of data easy and effective. By using a dedicated system of symbols, data elements are labelled and made comprehensible internationally.

---

4 Cf. https://access.rdatoolkit.org.
5 Cf. https://wiki.dnb.de/display/STAC/FG+Erschliessung.
6 Cf. https://wiki.dnb.de/display/STAC/STA-Community.
7 Cf. https://www.ifla.org/publications/international-standard-bibliographic-description.
8 Cf. https://www.ifla.org/.

JLIS.it

In recent years, the importance of the ISBD has waned slightly in Europe and North America. The standard is no longer in step with the times in terms of publication type (print-based publication or PDF) and also fails to take account of modern publication formats such as audiovisual media. Furthermore, it also doesn't take account of the IFLA Library Reference Model (IFLA LRM)[9] developed in recent years. However, a survey conducted by the IFLA has shown that this standard is still very widely used in some parts of the world where there is a complete (or partial) lack of stable infrastructure. Furthermore, the ISBD is regarded as easy to learn and apply, including by employees who don't have advanced professional qualifications. For this reason, the IFLA ISBD Review Group[10] decided two years ago to fundamentally revise and update the standard. Along with revising it in line with the IFLA LRM, it is being restructured and adapted to modern conditions. The basic principle of user-friendliness and the possibility of performing simple cataloguing tasks with it are to be retained, however. In addition to its future publication in a web-based environment, the standard will continue to be available as a PDF document and to print out. The initial work results of this update are expected in 2022.

Conclusion

Despite their many differences, all three of the aforementioned examples have certain things in common. They are all being created in a stable organisation culture. There is a committee taking responsibility for their development and revision, and supporting this work by providing resources. As different as they may be, all three standards focus on practical application and are geared towards simplicity and feasibility whilst simultaneously achieving the highest possible degree of standardisation. All three examples are being developed collaboratively and in direct communication with the respective user community. These commonalities seem to be a key factor in the success that unites these otherwise very different standards.

At the same time, these three approaches also highlight the fact that there can be no catch-all solution and that no single standard can ever adequately cover every practical application. This is even more true when we abandon discipline-specific approaches and start to think in more general and interdisciplinary terms. Every previous attempt to create a one-size-fits-all standard has failed. However, in this insight lies the future of standardisation within the realm of cultural heritage. Only modular, model-based frameworks will prove capable of ensuring the necessary flexibility and compatibility. Based on this fact, user communities must make adaptations in line with their needs that can be implemented in practice. In the long term, none of the cultural institutions will be able to employ a sufficient number of employees with the ability to implement highly theoretical standards. In light of the overwhelming amount of (digital) material that will need processing in future, this would also be a completely pointless endeavour. Keep it simple, but keep it standardised!

---

[9] Cf. https://www.ifla.org/publications/node/11412.

[10] Cf. https://www.ifla.org/isbd-rg.

JLIS.it

Bibliographic control in the fifth information age

Gordon Dunsire[a]

a) Independent Consultant, http://orcid.org/0000-0003-2352-0802

ABSTRACT

Bibliographic control is concerned with the description of persistent products of human discourse across all sensory modes. The history of recorded information is punctuated by technological inventions that have had an immediate and profound effect on human society. These inventions delimit five 'information ages'. It is now the Fifth Information Age, characterized by the ubiquitous use of powerful portable information processing devices for peer to peer communication across the entire planet. All such discourse is recorded during transmission and is copied to persistent storage media.

In the Fifth Information Age, the end-user is immersed in and interacts with a global ocean of recorded information. The interaction is continuous and ubiquitous, and never passive. Every interaction increases the volume of data; all aspects are recorded, including the time, place, and nature of the interaction, and details of the 'reader' and their 'book'. The roles of cave 'artist', scribe, printer, publisher, encoder, broadcaster, librarian, and other mediators are no longer differentiated from 'author'. The distinction between data and metadata is completely blurred: data becomes metadata as soon as an information resource is named by its creator.

The challenge for bibliographic control is the reconciliation of globalization and personalization via localization. The bibliographic ecosystem is very different and the activities and imploded roles of the end-user must be taken into account by professional agents.

KEYWORDS

Bibliographic control; Semantic Web; metadata; information retrieval.

JLIS.it

Paper

The context in which 'bibliographic control' takes place has been evolving at a fast pace for the past 30 years. Usage of the term was initially confined to written materials held in library collections, but has broadened to cover a wider range of information resources held in a wider range of collections. As a result, it is necessary to clarify the definition that is used in this paper.

The report of the Library of Congress Working Group on the Future of Bibliographic Control published in 2008 defines bibliographic control as "the organization of library materials to facilitate discovery, management, identification, and access" (Library of Congress Working Group on the Future of Bibliographic Control, 2008).

The IFLA Library Reference Model (LRM) published in 2017 is intended to cover "everything considered relevant to the bibliographic universe, which is the universe of discourse …" (Riva, Le Bœuf, and Žumer 2017, 20). The LRM is an entity-relationship model that consolidates three previous models for bibliographic records, authority data, and subject authority data published by the IFLA (International Federation of Library Associations and Institutions) as part of its development of "universal bibliographic control" (UBC). Although IFLA ceased its core support for UBC in 2003, development of bibliographic standards continues and the concept of UBC was "reaffirmed" as a set of principles in 2012 (IFLA, 2012). These principles are focused on the role of national bibliographic agencies and international coordination, and they include archives and museums in their scope.

The scope of the LRM is given by the definition of its broadest entity "Res": "Any entity in the universe of discourse" (ibid.). Dictionary definitions for the term 'discourse' emphasize written or spoken communication, and some specify a scholarly or formal context. For example, as of April 12, 2021, the online dictionary Dictionary.com gives two general definitions: "communication of thought by words; talk; conversation" and "a formal discussion of a subject in speech or writing, as a dissertation, treatise, sermon, etc." However, the LRM clearly intends a broader scope, beyond language-based materials, by giving examples of image, cartographic, and music resources. The LRM also restricts the definition to recorded communication: a resource is assumed to be embodied in a persistent carrier that can be accessed in the future, so speech must be recorded or transcribed if it is to be described.

The term 'bibliographic control' is defined by Dictionary.com in April 12, 2021 as "the identification, description, analysis, and classification of books and other materials of communication so that they may be effectively organized, stored, retrieved, and used when needed". No distinction is made between archive, library, and museum collections, and objects of control are "materials of communication".

This paper will therefore assume that bibliographic control includes all forms of recorded human communication. The 'bibliographic universe' is the set of all products of human discourse that forms the collective memory of Homo sapiens, and 'bibliographic control' is its management for future access and use.

Relevance

The bibliographic universe requires control because the organization of human memory is nec-

JLIS.it

essary for social cohesion and cultural evolution. Recorded discourse is communication through time and across distances greater than the unassisted range of human senses.

Recorded discourse carries the information that allows humans in different family groups to co-operate with each other in larger social units. The persistence and accumulation of recorded memory drives culture and its evolution. The inheritance of recorded memory is essential for cultural identity; the bibliographic universe is synonymous with cultural heritage. The management of recorded memory improves its utility and functionality in this context.

Recorded memory is an intermediary stage in the communication of a message from one person to another. The message is transmitted and then frozen in time; the message waits to be received at some unknown time in the future by some unknown person. The focus of bibliographic management is therefore the connection between the message and the receiver: what happens, after the memory is recorded, to the product that is recorded discourse?

The five laws of library science proposed by S.R. Ranganathan support this point of view (Ranganathan, 1931). The need for bibliographic control is driven by all five of the laws, although the terminology reflects a narrow focus on the written, and in particular printed, products that characterized libraries at the time. As of April 12, 2021, the Wikipedia article on "Five laws of library science" describes several subsequent attempts to modernize the scope of the laws and augment them to take account of the impact of more recent innovations in communication and information technologies.

The second and third laws are "Every reader his or her book" and "Every book its reader" respectively. The model is readily extended to all of recorded memory: the 'book' is the message, the recorded memory, the product of human discourse, and the 'reader' is the receiver of the message. The terms will be used in this paper with these general meanings.

The first and fourth laws are "Books are for use" and "Save the time of the reader" respectively. The primary factors affecting the delivery of the book to its reader – the recorded message to its recipient – are its portability, reproducibility, and findability. Portability determines if the book is taken to the reader, or the reader to the book. Reproducibility determines if the book can be accessed by more than one reader at a time. Findability determines if the book exists and how it is to be accessed by the reader. This last factor is the realm of bibliographic metadata: data about data, a book that describes other books so that readers can access their contents, the organization of the products of recorded memory.

The fifth law is "A library is a growing organism". The number of books increases over time. Recorded memory grows as time goes by.

Information ages

The ongoing evolution of human society and culture is punctuated from time to time by an innovation in communication technology that has a revolutionary impact. Such an innovation is followed by a significant increase in the complexity of interactions and activity across all social groups world-wide. Profound changes take place in commercial, legal, religious, and other cultural systems that affect all aspects of personal life.

Four specific innovations have had the greatest impact on the recording of human discourse. These are writing, printing, telecommunication, and the Internet.

JLIS.it

Each innovation provides a fundamental change in one or more of the basic aspects of preserving human memory and providing subsequent access to it. This results in a significant change in basic cultural and social concepts and processes; a paradigm shift. The innovation evolves through further invention and continues to influence many aspects of social interaction and development until the next innovation. It is useful to categorize the timespan between innovations as an 'age', and specifically as an 'information age'. The beginning and ending of each timespan are not precise dates, and they vary from place to place. Individuals and groups may recognize the potential for change that the innovation represents, but the actual impact of the innovation is not predictable during and immediately after the transition. Four innovations delimit five information ages; the present is the Fifth Information Age.

First Information Age

The First Information Age is the timespan before the invention of writing. It is pre-literate by definition, and is labelled "prehistoric" despite the existence of products of human discourse in the form of images and manufactured objects.

The production of a painting or sculpture takes time and requires specialist skills and tools, so such products are expensive. The fragility and perishability of available carrier materials means that only objects made of hard substances such as stone and images preserved under rare special conditions have survived. How widespread was the recording of human discourse is not knowable, but human groups were nomadic and small: Paleolithic and Mesolithic hunter-gatherers.

In this age, most social and cultural memory is conveyed into the future, beyond the individual memory of a person, through an oral tradition that cannot be recorded (until the invention of writing).

The content of the discourse that is recorded is mostly representational, depicting the things of interest in the local environment. Some content is symbolic and abstract, but the context is unknown. The meaning or intention of recording the content cannot be determined; only the 'art' can be appreciated in the context of modern aesthetics.

Reproduction of the recorded memory is as expensive as manufacturing the original. Each carrier of the content is a one-off, a singleton manifestation in the terminology of the LRM.

Access to recorded discourse is very limited. Images carried by cave paintings are often located in the furthest reaches of the cave. The reader must be taken to such a book to access it, and this seems to have been a religious or ritualistic activity. Portable sculptures must be small and light enough to be transported along with the other possessions of hunter-gatherer social units. Fragile carriers such as wood and soft stone are easily destroyed, small objects are easily lost, and such books are very rare. What has survived is now curated in museum collections.

Second Information Age

The Second Information Age begins with the invention of writing, the symbolic representation of language. Writing allows the recording of linguistic discourse. The act of speaking is readily transferred to the acts of writing and reading. The recording of discourse in specific aspects of human culture becomes common-place.

JLIS.it

The content of recorded linguistic discourse is descriptive and much more expressive than images and objects. There is immediate benefit in recording the 'word' in commercial, legal, and religious systems; social agreement is no longer reliant on the oral tradition or individual human memory. Peer-to-peer communication over long distances between persons who are known to each other, the writing of letters, becomes possible.

In this age, carriers remain singletons, such as manuscripts and paintings, but reproduction requires only the skills of the scribe or copyist. Reproduction has the same costs as the manufacture of the original manuscript, but this is less expensive than copying a painting or object. The process of reproduction is industrialized with the development of the scriptorium. Centralization of reproduction leads to centralization of storage, and the first libraries appear.

Access to recorded memory becomes easier. Readers who can travel independently can go to the scriptorium or library. Writing is applied to flat surfaces, and the third dimension of the cave or figurine is not required. This allows and encourages portability by embodying the message in materials such as clay, bark, bone, and textiles. Some writing is monumental, such as the Code of Hammurabi stele, and the reader must go to the book, but many products of discourse can be carried by hand to the reader. Not many survive because of the perishability of portable carriers.

Third Information Age

The Third Information Age begins with the mechanization of printing. Printing is a development of the industrialization of writing that involves the mechanical reproduction of writing and images. Development of the technology begins in the Second Information Age with the use of seals for stamping text onto clay or paper. The content is usually a name that confers ownership or authority on an accompanying manuscript. The technique evolves to cover the content of a page of text or a drawing in a larger stamp made of wood, stone, or some other hard material that can be sculpted. This speeds up the production of copies of texts and images, but preparing a seal or stamp is expensive and the range of discourse that is recorded in this way remains very limited.

The Second Information Age ends with the development of movable type and printing presses which industrialize the mechanics of reproduction. Manufacture and reproduction of the products of discourse becomes much less expensive, and there is a corresponding increase in the quantity of such products. Reproduction becomes part of the process, and the existence of multiple identical copies becomes the norm. The products of recorded discourse become more common-place, but are mediated by the printer who has the skills to set the type and operate the press.

There is an immediate and significant increase in the range of persons whose memory is recorded. A greater proportion of depictive content is manufactured and distributed using the new technologies, to cater for readers who are illiterate or who do not understand the language of a text; a picture bridges linguistic barriers. Scholarly communication becomes industrialized with the development of printed journals.

Access becomes easier. The reader has a choice of copies of the book, located in multiple places, and the book is easy to transport. Printers and booksellers become 'high street' services, and modern libraries begin to develop.

JLIS.it

Fourth Information Age

The Fourth Information Age begins with the invention of digital telecommunication. The development of the transmission of information over large distances required new techniques for correcting signal errors while increasing the size of the message; this stimulated the evolution of digital technologies.

Most forms of telecommunication require the message to be encoded so that it can be transmitted. The message is decoded back into its original form when it is received. The application of telecommunication technologies to discourse usually requires the discourse to be recorded as part of the encoding and decoding processes.

Encoding allows all forms of content to be transmitted, including music, speech, and static and moving images. In this age, the range and quantity of recorded discourse increases again. Electromagnetic media become available for the persistent storage of memory. Digital encoding allows the content and carrier of the book to be created, manufactured, distributed, and accessed in an integrated, seamless, and intangible infrastructure. Reproduction is unavoidable and invisible; a temporary copy of the product of discourse is automatically created in every encode/decode transaction and it is trivial to make that copy persistent.

There are no physical barriers to access, and access becomes localized; the book always goes to the reader, wherever the book and the reader may be. Transportation is instantaneous; the reader gets the book when and where the reader wants it.

Fifth Information Age

The Fifth Information Age begins with the invention of the Internet. The Internet globalizes digital telecommunication networks linked to powerful data processing machines and allows the participation of nearly every living human in discourse over a distance.

Digital encoding and decoding are a necessary process for discourse using the Internet. All discourse is recorded on persistent digital media. The deletion of recorded memory, "the right to be forgotten" (ICO, n.d.), has become a cultural and social issue, in a complete reversal of the First Information Age and 'the right to remember'. An example of the impact on bibliographic control is the initiative by NISO on "author name changes" (NISO, 2021)

The World-Wide Web is an application of the Internet that allows any person to take on and combine the roles of author, publisher, printer, distributor, and reader. The book includes every email, social media post, chat or webinar conversation, blog, website, or search ever made by every reader.

Reproduction is a built-in automatic feature. Overt reproductions of recorded memory are made to ensure persistence of cultural heritage, improve access, and retain evidence of discourse.

The "Internet of things" is a result of the miniaturization of computer chips as digital encoding, storage, and decoding devices. The reader and the book exist in the same local space and time. The perceived benefits of allowing 'all cookies' ensures that recording is ubiquitous and constant; the 'user' is immersed in an ocean of recorded/recording memory. The reader is every individual human; the book is a collection of all digital human memory.

JLIS.it

Metadata

The development of metadata for bibliographic control arises in the Third Information Age. The quantity and availability of printed products stimulated an increase in collections of recorded memory by social groups and individuals. Such collecting began in the Second Information Age with the development of libraries of manuscripts, but these were rare because of the expense of obtaining or reproducing hand-made products. Printing allowed wealthy individuals to accumulate private collections for pleasure, research, and status, and for a greater range of commercial, legal, religious, and scholarly organizations to develop repositories of information to support their activities. As collections grew in number and size, it became useful to record the collector's memory of what the collection contained, and to organize access to the collection to find and select a specific product of discourse. Is the item in the collection, and if so, where is it located? "As the number of books available to collectors like [Hernando Colón] grew, and new ways of organizing them became necessary, a list of authors in alphabetical order probably seemed a fairly unproblematic place to start … the alphabetical list forces the librarian, and the users of the library, to attribute each of the books to a single, named author, in a sense 'inventing' the notion of the author (or at least their centrality) as a matter of necessity" (Wilson-Lee, 2018, 209-210).

The content of metadata is essentially descriptive, and therefore linguistic in form. Textual metadata can be sorted and ordered using the syntax of the language of description, and it is much easier to formulate search and retrieval queries in the same syntax. Textual metadata can be transformed into spoken word, using a screen-reader, or visual symbols such as colour-coded categorizations. On the other hand, depictive metadata content is of limited utility. A thumbnail image is a representation or depiction of the whole image, not a description of it. Essentially, the reader reads a (metadata) book in order to find a (data) book.

The Third Information Age therefore stimulated and supported the printing of metadata as a result of the printing of books. The Fourth Information Age stimulated the internationalization of metadata creation, reproduction, and distribution. The MARC formats were initially developed to be "a vehicle for the exchange of bibliographic information between systems with independent computer facilities" (Morton, 1986). The Fifth Information Age allows the reader to be the author and publisher of metadata – the cataloguer – as well as being the author and publisher of a book that is being described.

Current approaches to metadata are rooted in the paradigms of the Third and Fourth Information Ages. The impact of the Fifth Information Age on bibliographic control is at its beginning and the detail belongs to the unknown future, but it will be profound. Some of the main characteristics of the bibliographic future are already emerging, including identity management, data provenance, open world application, and the authenticity of consensus.

Identity management

The management of identity is essential to the functionality of metadata. An identifier is a label that distinguishes the referent from other things. Effective information retrieval processes require that the subject of a metadata description is identified: is the individual book or associated entity that is being described the one that the reader wants?

JLIS.it

Identity management is the basis of classical authority control, a development of the concept of 'author' from the Third Information Age. The nature of discourse, and human culture itself, differentiates names and titles in specific social contexts only; there is no global system that makes the distinction based on universal physical contexts such as space and time. A person is not a cultural artefact, but is a natural phenomenon that cannot exist in two places at the same time. The same person has different names; the same name can refer to multiple persons. This is surely a result of larger, settled groups in the Second Information Age. More generally, the same individual is labelled with different identifiers, and the same identifier is used for different referents, across different human cultures. Much of this diversity is driven by local context and by the difficulties of assigning identifiers that are agreed at global level.

The Fourth Information Age stimulated the development of global approaches to identifier management, generally limited to the book and its trade. Examples include the International Standard Bibliographic Number (ISBN) and International Standard Serial Number (ISSN) systems. The beginning of the Fifth Information Age saw the development of similar approaches to the identities of persons, including the author and therefore ultimately the reader, such as the International Standard Name Identifier (ISNI) and ORCID. However, it is not always a single person or group of persons that is being identified, and the cultural confusion of names and named persists, as ISNI's name suggests. As of April 12, 2021, the ISNI website states that it covers "public personas … such as pseudonyms, stage names, record labels or publishing imprints"; the LC/NACO Name Authority File remains under active development. The LRM includes an entity Nomen, the class of names of things, that is distinct from the things, such as agents, places, and timespans, themselves. This allows description of the name, such as usage, language, etc. to be separated from description of the thing that is named.

However, the Fifth Information Age eliminates half of the general problem, of the same identifier being used for different referents. The Internationalized Resource Identifier (IRI) system, based on the Uniform Resource Identifier (URI), is applicable to anything that can be described; that is, any thing that is the subject of bibliographic metadata. This is one of the necessary and fundamental aspects of the Internet, the World-Wide Web, and the linked open data of the Semantic Web. It is managed independently of any cultural application or context.

The assignment of more than one identifier to an individual thing cannot yet be eliminated. That would require all of the assigners of identifiers to agree on a preferred identifier and to supply a means of de-referencing it to a description of the thing it identifies. This was the approach of IFLA's UBC programme, and is the antithesis of the bottom-up construction of the Semantic Web. In the Fifth Information Age, authority control evolves into the management of linked data identifiers. The application of automated reasoning to connect the reader to the book is completely dependent on consistent and complete assignment of IRIs to readers, books, and associated entities. It is important that there is no ambiguity in what is being identified within the chosen data model, such as the LRM or BIBFRAME. The rules used in semantic reasoning are simple and they are applied by dumb machines; it is the metadata that is 'smart'.

Data provenance

The Semantic Web is a globalized metadata retrieval system built on the World-Wide Web. It is based on description logic and has no intrinsic accommodation of "truth". The Semantic Web adheres to the AAA Principle: "anybody can say anything about any thing"; this is alternatively known as the AAA Slogan: "anybody can say anything about any topic" (Allemang and Hendler, 2011, 27). What is said in metadata may be true or false, in the same way that the content of any product of discourse may be true or false relative to the context in which it was created. Statements may be true when recorded, but are false when they are replayed; things change. Statements may be known to be false when recorded. "This statement is true" may be fake, and its author a liar. This is not just a cultural phenomenon. Discourse itself has in-built paradox, ranging from the "impossible" images of M.C. Escher to the linguistic paradox of Epimenides: "This statement is false" is false if it is true, and true if it is false.

These uncertainties mean that effective bibliographic control requires provenance for metadata. This is metadata that describes metadata, and has similar functionality to data provenance or "detailed information about the origin of data" (Glavic and Dittrich, 2007). For bibliographic metadata, provenance includes information about the author (cataloguer, curator, etc.), the application of content and encoding standards, and the date of creation. Data provenance has been accommodated in bibliographic control from the Fourth Information Age to support the coordination of shared catalogue records. For example, this is provided by leader and control fields in MARC formats, such as "Date and Time of Latest Transaction" (Library of Congress Network Development and MARC Standards Office, 1999). Another latent example is the use of brackets in International Standard Bibliographic Description (ISBD): "Square brackets enclose information found outside the prescribed sources of information and interpolations in the description" (ISBD Review Group, 2011, 22). The recording of bibliographic data provenance for more general purposes is given specific accommodation in the development of more recent standards such as RDA: Resource Description and Access (RSC Technical Working Group, 2016).

Provenance is a means of quality control. Knowing who created metadata helps to distinguish high-quality data created by trained professionals with ethics from low-quality data created by amateurs with bias. It is also important to know when metadata was created and what standards were used. Metadata theory and practice evolve just as much as any other form of discourse. How things were described in the past may be useless or misleading in a contemporary context. Provenance allows metadata from disparate sources to be aggregated without 'one bad apple' lowering the quality overall.

Open world

The Semantic Web also makes the Open World Assumption (OWA). The assumption is that the absence of metadata is not a description of absence, but simply a description that has not yet been made. Metadata may be added in the future, and there is no expectation that future metadata will be objectively or subjectively true. This is a consequence of the AAA principle and the paradox of discourse: there cannot be a complete description of a thing because an infinite number of false or unprovable statements can be added.

Applications of bibliographic metadata based on closed-world assumptions become less efficient in the Fifth Information Age. A bibliographic record can no longer be a fixed and complete description of a book or the entities associated with it. Metadata will always accumulate, so the size of the 'record' increases through time. It is unlikely that any single application will need or want to use the whole set of metadata that describes an entity, but the set exists and cannot be ignored. The closed-world practice of updating erroneous or incomplete metadata is no longer tenable. Instead, it must be assumed that the original statement of metadata is 'out in the field' in multiple information retrieval systems where it is not feasible to update every copy. Revisions are made with new statements; erroneous statements are assigned appropriate data provenance.

Wikis that share data from multiple authors without central mediation have been involved in conflicts where statements are updated by one author and 'updated' back to the original statement by another author. Each author wants their version to be published and the other's version to be discarded. For example, Wikipedia has a published policy on "dispute resolution" that seeks consensus before arbitration is invoked. As a result, data provenance and version control systems built-in to wiki software have become an important tool in quality control and assurance. Nothing can be truly deleted in a wiki, and amendments can be 'rolled-back' to a previous version. Similar systems are required for metadata.

Imposing fees for the use of metadata in wide-area applications or for the copying of metadata to use in local applications is a barrier to the utility of metadata in the Fifth Information Age. It prevents open linking and discourages the reader's contribution of metadata to the global pool, for example through passive cookies or active crowd-sourcing.

Consensus

If any reader can make any metadata statement they want, with no distinction between 'fact' and 'fiction', how can any consistency or authenticity be determined?

In the Fifth Information Age, recorded discourse is cultural memory, and metadata is the organization of culture itself. What makes local culture consistent is local consensus. A social group agrees to a particular set of truths, reflected in its recorded memory, to maintain a consistent and persistent world view.

Consensus in metadata can be determined through analysis by machine and by the human mind. Statistical analysis of large sets of metadata accumulated from multiple sources can calculate consensus by matching similar statements and by using data provenance to detect bias from particular sources. This is basically how search engines work; relevance is determined by the automatic analysis of the links on a webpage, where the focus of the page is assumed to be the subject of the link, and the links to a webpage, where the page is the target of the link. The link itself is metadata; the subject and target are associated in some way.

Linked open data in the Semantic Web can be processed using semantic reasoning, a standard set of algorithms that can derive metadata statements from metadata statements. These algorithms are simple, reflecting the simple 'atomic' structure of the linked data subject-predicate-object triple. They are not a substitute for human intelligence and culture. These automated techniques are a tool for cataloguers, not a substitute for cataloguers or other humans.

JLIS.it

Human analysis of metadata may be conscious or subconscious. The reader carries out such analysis throughout their information seeking and retrieval activity. The conscious analysis of the relevance of data is a form of 'ask the audience' in a quiz show. This is a core feature of social media in the Fifth Information Age, where the audience is invited to like or dislike (choose a binary review of) a piece of data, a mini-book. Consensus is reflected in the numbers of persons who like or dislike the information and the balance between them. This is a very broad measure of the 'authenticity' of data or metadata. A more refined approach is to crowd-source contributions for specific sets of books by specific sets of readers.

Subconscious analysis is now possible using eye-tracking technologies. The reader has no control of how their eyes read a book or description of a book. Experiments show that it is not the linear scan that it appears to be in the conscious mind. The development of virtual reality, mimicking the immersive cultural memory of the Fifth Information Age, will stimulate the use of subconscious feedback technologies.

Effectively 'author', 'authority', and 'authenticity' blur into the control of culture by consensus.

Conclusion

The future of bibliographic control is as unpredictable as the future of writing, printing, telecommunication, or the Internet when they first appeared. In every case, there has been an immediate impact on human discourse and recorded memory, followed by a slower but profound impact on every aspect of human culture. Although the dates may be imprecise and localized, the timespan of each information age decreases by at least an order of (decimal) magnitude, from tens of thousands of years through a few thousand and a few hundred years to a few decades.

Syntactically rooted in the Second Information Age, conceptually rooted in the Third Information Age, and mechanically rooted in the Fourth Information Age, bibliographic control is struggling in the Fifth Information Age. The range and quantity of products of recorded discourse requires a shift in the focus of bibliographic control, from top-down to bottom-up with the 'professional' cataloguer distinguished from other readers by context, not process.

Bibliographic control is likely to be based on the Open World Assumption. It will involve the coordination of metadata created by professionals and amateurs with metadata created by machine analysis. Data provenance is essential to achieve this by providing context and supporting the management of quality control. Metadata is common and necessary in the Fifth Information Age. It is a social and cultural 'good' that is not best controlled by commercial interests.

The purpose and function of bibliographic control is to manage cultural identity in a global framework. The distinction between data and metadata is no longer useful, and bibliographic control will become indistinguishable from culture. The Fifth Information Age is the technological extension and immersion of personal and social mind.

JLIS.it

References

Allemang, Dean, and Jim Hendler. 2011. *Semantic Web for the Working Ontologist,* Second Edition. Amsterdam, Morgan Kaufmann.

Glavic, Boris and Klaus R. Dittrich. 2007. "Data Provenance: A Categorization of Existing Approaches" In 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme", Aachen, Germany, 7 March 2007 - 9 March 2007:227-241. Accessed April 12, 2021. http://dx.doi.org/10.5167/uzh-24450

ICO. nd. "Right to erasure". Accessed April 12, 2021. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-erasure/

IFLA. 2012. *IFLA Professional Statement on Universal Bibliographic Control.* Accessed April 12, 2021. https://www.ifla.org/files/assets/bibliography/Documents/ifla-professional-statement-on-ubc-en.pdf

ISBD Review Group. 2011. *ISBD : International standard bibliographic description, consolidated edition.* Berlin, De Gruyter Saur.

Library of Congress Network Development and MARC Standards Office. 1999. *MARC 21 Format for Bibliographic Data: 005 - Date and Time of Latest Transaction (NR).* Accessed April 12, 2021. https://www.loc.gov/marc/bibliographic/bd005.html

Library of Congress Working Group on the Future of Bibliographic Control. 2008. *On the Record: Report of The Library of Congress Working Group on the Future of Bibliographic Control.* Washington, D.C.: Library of Congress. Accessed April 12, 2021. https://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

Morton, Katherine D. 1986. "The MARC Formats: An Overview" In *American Archivist* Vol. 49, No. I/Winter 1986:21-30.

NISO. 2021. "NISO Members Approve Proposal for a New Recommended Practice to Update Author Name Changes". Accessed April 12, 2021. http://www.niso.org/press-releases/2021/04/niso-members-approve-proposal-new-recommended-practice-update-author-name

Ranganathan, S. R. 1931. *The Five Laws of Library Science, etc.* Madras, Madras Library Association.

Riva, Pat, Patrick Le Bœuf, and Maja Žumer. 2017. *IFLA Library Reference Model: a Conceptual Model for Bibliographic Information.* Den Haag, IFLA. Accessed April 12, 2021. https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf

RSC Technical Working Group. 2016. *RDA models for provenance data.* Accessed April 12, 2021. http://www.rda-rsc.org/sites/all/files/RSC-TechnicalWG-1.pdf

Wilson-Lee, Edward. 2018. *The Catalogue of Shipwrecked Books: Young Columbus and the Quest for a Universal Library.* London, William Collins.

JLIS.it

Follow me to the library!
Bibliographic data in a discovery driven world

Richard Wallis[a]

a) Data Liberate, http://orcid.org/0000-0001-8099-5359

ABSTRACT

Libraries are generally welcoming organisations and places. Engaging with communities, inviting all comers to immerse themselves in the information rich environment curated for the benefit of all, from the entertainment seeker to the educational specialist. Traditionally this immersion would take place in open welcoming impressive buildings at the heart of the town square or university campus.

However, as witnessed by the phenomena of the declining town centre and the lockdown Zoom culture of 2020, traditional routes to resources are changing rapidly. In the online discovery and delivery world that has emerged, metadata especially quality metadata, about resources and information is key. Without a detailed understanding of available resources, it can be difficult if not impossible to direct them towards those that might benefit from reading, watching, analysing, interacting with, or purchasing them.

JLIS.it

Hello everybody. Thank you for the organizers of this conference and for inviting me. I hope you'll find this interesting. So here we go. First, if you've never met me. I'm an independent consultant. I have been around computing, far too long to own up to, but involved with cultural heritage technology for a significant period of time and with the Semantic Web and Linked Data since they were first introduced.

I have been involved in the W3C consortiums heading up community groups mostly around bibliographic, archive, financial data etc., and the standard schema.org – of which I'll come on to in a minute.

I work with various organizations. I work with Google (not for Google) helping them to contribute to the open schema.org[1] vocabulary project. I have a lot of involvement: making sure the site still runs in that area, extensions, documentation, community engagement etc.

I have worked with my previous employers (OCLC), financial industry people and with various clients that are relevant to this conversation. I've worked with The British Library, Stanford University Law, Europeana, and National Library board of Singapore.

The reason I am here is to talk about the way we may have to change our approach.
I am using the analogy of libraries all the way through this conversation, but it could equally be archives, it could be museums, it could be aggregators like we heard about it in the previous presentation.

Libraries have a reputation of being welcoming places usually in settings of imposing or inviting buildings in town and city centres or having an imposing but important place on university campuses. Within the buildings offering the right sort of environment for people to read and study etc. That includes more social spaces, often found on university campuses. We reach out to plus possible users at an early stage inviting schools into libraries etc. Once people are into a library. Once through the door it becomes a little intimidating when you first come in, because your first challenge is to find stuff. Traditionally this was done within impressive wooden sets of drawers with catalogue cards in. Those catalogue cards evolved into some standard formats that were used – often a little obscure to the users – but the librarians were usually quite pleased with these.

When technology turned up that could help us, in the 1960s, libraries adopted it very rapidly; which led to the arrival of the MARC record card this introduced cataloguing data standardization.
We needed standardization so that the computers could work across the data and understand it; and let us build some systems; and those systems enabled us to roll out the catalogue for people to search and interact with, it well beyond those little wooden drawers across the library and very often into the outside world.

---

[1] https://schema.org/

JLIS.it

Talking about the outside world…
During this period our world has changed. You only have to interact in the outside world, and you find that we're moving away from the traditional town centre, or city centre, and moving towards shopping destinations or entertainment destinations etc.

We saw the growth of out-of-town shopping centres, which introduce efficiencies for the retailers and a destination for the shoppers. This has evolved even further to online retail, which has delivered massive efficiencies for the retailers, and kind of removed the environment that the library used to interact with and led to what people have christened the death of town centres. The inevitable move to an online culture has been what we've been readily aware over the last 12 months. We have moved into the Zoom society, where people interact for business meetings or family occasions etc. This has been exacerbated by recent lockdowns.

Libraries started to react to this move by starting to reach out even more, and become attractive – be it in the public arena or in the academic arena – providing social spaces and traditional non-library spaces (often becoming as concerned about the quality of the coffee as the quality of the reading materials). This translated online as well. So, the initial computerized catalogues were fairly dry affairs that emulated the original card catalogue. We started to add book jackets and links to other resources. Following on with, what traditionally got called web 2.0
Standards, where it became far more graphical and started to emulate the online retailers. Not only in the look, but in the searching capability. Moving away from the traditional keyword within title searching that was available, towards entity-based searching. So, searching often returns now sets of authors or organizations or works or articles or whatever.

Equally following the evolution of the look and feel, of the rest of the world, it started to move in some libraries – this is the public library of Oslo[2] – to a much cleaner much more graphical environment. As our potential users started to understand these environments, because they're using them all day every day for their social networking and other things, and if you take this sort of approach you can start moving into a very visual environment that would operate on a an iPad or a mobile phone of some sort.

What's going on behind the scenes to enable all this?
There was a set of developments where we started to inherit that the technology for the rest of the world. The move from MARC to MARCXML enabled us to use standard programming tools to start working with bibliographic data. Starting to use in some circumstances simplified vocabularies to describe. We introduced textual indexes to enable the interpretation of the material beyond what were the capability of relational databases, and with the introduction of Linked Data RDF (the Resource Description Framework) started to enable us to move in a Linked Data direction.
The introduction of BIBFRAME[3], from the Library of Congress, in about 2011-2012 which was

---

[2] https://deichman.no/sok/abbey
[3] https://www.loc.gov/bibframe/

JLIS.it

the first approach of the very detailed bibliographic vocabulary to describe bibliographic resources as linked data.

RDA picked upon the linked data theme and started to introduce a Linked Data version of RDA along the way and, more recently, BIBFRAME 2.0 came out which reflected some of the challenges that were encountered with the initial BIBFRAME, making it easier to operate within a linked data world.

So that's what's been going on in the library world. What's been going on in the wider world?...
Well, the wider world has been taking on something that they call Structured Data.
Why are they doing that?...
Well, there's two driving forces for it. There's the search engines, Google and their colleagues, who have come to the end if you like of the capability of being able to confidently mine textual data on the web page, to work out what the thing was and its attributes that the page was describing. Equally the publishers of websites had hit the point where they wanted to be
able to more accurately describe their resources, or things, to the search engines.
Early attempts included: calendar and business card formats, that could be embedded in the page, Google had an attempt at an open vocabulary called data-vocabulary.org. Eventually in 2011 Schema.org arrived on the scene. Backed as an open project by Google, Bing, Yahoo! and Yandex. This introduced the standard that has since been taken up by others like Facebook, Alexa, Apple, Pinterest etc.

So, what is Schema.org?...
It's an open vocabulary for the web, a Linked Data vocabulary (although though they don't shout about it too much), RDF based. It's got well over 2000 terms in it (778 types, 1383 properties in the last release). It releases every two or three months, so it evolves. Basically it means, in the wider world, most things have got a type, in Schema.org, that can be used. Things like creative works (CreativeWork), persons (Person), volcanos (Volcano), libraries (Library), medical procedures (MedicalProcedure), books (Book), etc. And this wide vocabulary and ease of use has enabled it to deliver a significant penetration
There is an open crawl project that happens every year, in 2020, round about September time, they did a crawl and they crawled 3.4 billion urls that were held on 34 million domains – so quite a significant chunk of the web. They identified that 44% of those domains had structured data embedded on them (mostly Schema.org), and 50% of the pages.[4] So 50% of three point four billion pages, had Schema.org or similar embedded on them.

The key is it's embedded in the HTML so the publishers don't have to do anything technically clever. They don't need specific endpoints to query the data. They just embed it in their normal HTML website in one of three formats to choose from (Microdata, RDFa, and most popular nowadays the JSON-LD)[5].

It gives you visibility on the web. What does that mean?...

---

[4] http://webdatacommons.org/structureddata/2020-12/stats/stats.html
[5] https://en.wikipedia.org/wiki/JSON-LD

JLIS.it

On a search engine, you can obtain rich results or be part of a knowledge panel. This screenshot from Google (see Figure 1) is showing you that the bottom but one result is a rich result so it includes things like ratings and pricing information. They often include an image etc whereas the one below it for the same item is just a boring ordinary listing. Whereas in the knowledge panel, the data harvested from all sites describing this thing is an aggregate representation of what that 'thing' is about and what it's related to.

It also drives specialist services. I'm only going to pick up one now because we haven't got a lot of time. That is Google's Dataset Search (see Figure 1), which is a specific search and looking for data sets that are openly shared on the web:[6] You see a lot of Covid-19 data sets are being shared at the moment, academic data sets etc. The key to this is, unless you embed Schema.org in the page that describes that data set, and where to get it, you almost certainly won't end up in Dataset Search.
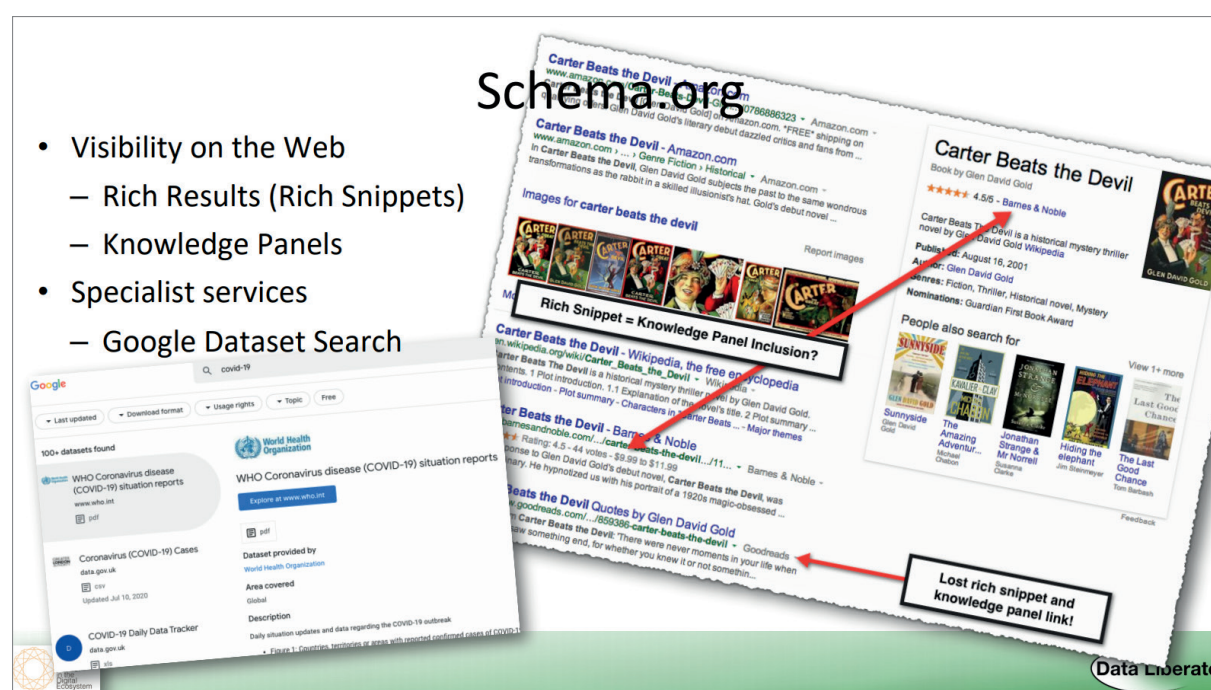
Fig. 1.

So, what's going on under the hood on the websites that are doing this?...

They want to describe their 'things', they want to describe their products, their events, their services, offers, articles, persons and organizations – that are for sale or to lend or whatever. To do that they have to mine their data. Quite often that's done by just updating the software that builds the HTML page to encode the data, that the page is about, inside the HTML. Or sometimes it's extracted from databases, and APIs are used to create a Schema.org structured payload that gets embedded in in the page. The search engines and others – it's open for anybody to crawl – extract the HTML from the pages, and from within that, extract the Schema.org structured data for use.

---

6 https://datasetsearch.research.google.com/

JLIS.it

So, we have got different data practices:

- *libraries* use Linked data -- the *web* use Linked data;
- *libraries* use detailed standard vocabularies (RDA,BIBFRAME, etc) - whereas the *web* is using a common global general purpose vocabulary, schema.org;
- *libraries* quite often use this structure [Linked data] to make it easier for them to link often externally -- the *web* is, almost by definition, totally externally linked (even within your own website the linking is identical) So whether you're linking to a Wikipedia article or another page on your own site it's the same principle. [an entity linking oriented];
- the *libraries* have used this to start delivering enhanced discovery service interfaces, entity based local searching (such as I exampled earlier on), improved detailed display (so it's there to improve the discovery interface) -- whereas on the *web* output in schema.org gives in enhanced data for search engines, rich results display, representation in knowledge graphs. Which almost by definition means you're far more likely to receive accurate links from the search engines into your resources.
- for *libraries*, the standards and the uses are for libraries and partner libraries only (or things like aggregated catalogs which we saw earlier on) -- whereas out on the *web* the structured data is for growing global representation and linking.

So, what are libraries doing on the web at the moment?...
Well mostly (there are some exceptions), basically the web knows about your discovery interface (hopefully) or (more likely) the homepage of your website. Not the things you can discover using that interface. Users do not start their discovery journey in your interface, they're not looking for you (the library), they're looking for the resources that you can provide.

So how do we get our resources visible in the web?...
The answer is quite simple: start sharing Schema.org data from our discovery interfaces.

The answer might be simple, the implementation might be a little bit more of a challenge but more of that in a moment. Let me show you an example of bibliographic data on the web[7]. This is the National Library Board of Singapore (I have worked with them for many years). Here what they have done is, taken every catalogue record from their library system, and produced a static web page that describes it – very, very catalogue card-ish I would suggest. They've enhanced it fractionally by adding an image (if they've got one), and a link to the library system it came from. There is no search interface for this. There are just thousands and thousands of static web pages on a website. But, embedded in those pages is Schema.org. This is the structure that's in there describing the entity, and if you want to actually look at the JSON-LD that's embedded in the page – if you're that way inclined this is what that JSON-LD looks like[8].

---

[7] https://www.nlb.gov.sg/biblio/12343857
[8] Example from https://schema.org/Book

```
<script type="application/ld+json">
{
  "@context":  "https://schema.org/",
  "@id": "#record",
  "@type": "Book",
  "additionalType": "Product",
  "name": "Le concerto",
  "author": "Ferchault, Guy",
  "offers":{
      "@type": "Offer",
      "availability": "https://schema.org/InStock",
      "serialNumber": "CONC91000937",
      "sku": "780 R2",
      "offeredBy": {
          "@type": "Library",
          "@id": "http://library.anytown.gov.uk",
          "name": "Anytown City Library"
      },
      "businessFunction": "http://purl.org/goodrelations/v1#LeaseOut",
      "itemOffered": "#record"
    }
}
</script>
```

The search engines are pinged to say these pages are available. They provide a sitemap[9] to tell the search engine where to crawl and then they leave it to the search engine.

So, what's the effect?...
Well here we're seeing the effect. This is a snapshot out of the Google search console which is reporting traffic to that site. It's a 28 day period and in that period 1.58 million times one of those pages appeared in a set of search results. 61,000 times somebody clicked from one of those search results through to the site, and many of them clicked on to find in the library etc.
So that's a 3.9% click-through rate which – if you speak to any SEO expert – is not bad actually, especially for a static page – that hasn't got any kitten videos or similar.
So, this is something I believe most libraries would like to do. But this delivers a bit of a dichotomy if we use BIBFRAME.

BIBFRAME is a library standard, it's replacement for MARC, it's led by the Library of Congress, system suppliers are investing in it (at least importing or exporting BIBFRAME), it benefits the local interface, and libraries implementing this kind of thing see it's a significant step forward. A step forward, that is a development goal for many libraries and aggregators. So that you could almost do a very similar list, about most of the new technologies that are being worked on in the library world.

Whereas Schema.org is a global standard, it's not a library standard, it's backed by the search engines (and others of course), library suppliers are kind of looking at it but not really investing heavily, it benefits the global discovery and linking of your resources (not the local interface). It's a different step forward, and at best appears to be on the agenda in most libraries as a 'nice to have'.

---

9 https://www.sitemaps.org/it/protocol.html

JLIS.it

So, when I'm talking to libraries about trying to attempt to do what I am describing here I get answers like: "*well, BIBFRAME is taking our current focus*", "*schema.org is a different data model*"; "*we can't do both*" – well maybe we can.

As a world we're investing in linked data, it's the subject of many, many presentations on conferences like this and BIBFRAME tends to be the default Linked data standard for sharing your library data (there are others, don't shout at me in the chat).

We can build on this investment: not replace it but add in something like Schema.org on the back end of it. This is the subject of a W3C community group entitled Bibframe2schema.org which I chair. The objective is the creation of a reference mapping from BIBFRAME 2.0 to Schema.org, and the development and sharing of reference software implementations for people to copy. This site is very small at the moment but, if you look on the comparison viewer it will bring in BIB-FRAME 2.0 records and demonstrate an initial prototype conversion to Schema.org – so that we can see what the effect would be. So if we could reproduce this, if we could produce Schema.org into our discovery interfaces, so that the search engines and others can crawl it; we have the digital way to share digital breadcrumbs across the web, to draw people to our resources.

They don't have to find us first, and then learn how to use our specific interface. Their day-to-day tools, their questions to Alexa etc, should be able to pick up these breadcrumbs wherever they may be. To deliver the value of your resources, that you're spending a lot of time in an effort in encoding, and building standards around them and sharing in your own interfaces. Most of your users want to be able to get that at them from where they get up in the morning if you like. So, to be visible on the web we need to get our internal Linked data right.

BIBFRAME is a good candidate for this (not the only one). But we mustn't expect the rest of the world to use our vocabularies. Having a fully Linked Data catalogue is not going to do a lot, for people finding your resources across the web. Outputing the global de facto standard vocabulary Schema.org, *as well as* our relevant detailed vocabularies make this, as a community, easy and consistent for developers and implementers.

So, as the BIBFRAME world implemented MARC2Bibframe, which is a piece of software that you can use which will take a MARC21 record and produce Bibframe 2.0 data; equally we should be able to take Bibframe2Schema.org outputs and produce software that will take, the already produced BIBFRAME data and translate it into Schema.org terms which we can then fairly simply embedded in our user interfaces.

And to make this work we need, as a community, to participate in the community groups, participate in the discussions – participate in the web so our users can find the resources that we actually have on the shelves, and on our disk drives, for them.

Thank you very much.

JLIS.it

Collocation and Hubs.
Fundamental and New Version

Sally H. McCallum[a]

a) Library of Congress, http://orcid.org/0000-0002-6137-2129

ABSTRACT

This paper discusses collocation as a fundamental concept of metadata description that is reinterpreted and expanded in the BIBFRAME library linked data environment via the development of "hubs". With the MARC title authority description as a basis, the relationships that support broader collocation are examined and the affinity of the MARC title authority to a bibliographic entity is explained. The reinterpretation of the title authority as a bibliographic hub will assist the fluidity needed in today's environment between the MARC format, used for the last 50 years, and the new BIBFRAME ontology intended to replace it for richer linked data applications.

KEYWORDS

Metadata; BIBFRAME; Library linked data; MARC.

JLIS.it

Collocation

Collocation of information items has been a primary purpose of rules for bibliographic descriptions for a very long time. It was stated by Cutter in 1889 (Cutter 1889), well-articulated by Seymour Lubetzsky in the 1950s, and then reaffirmed and refined by the Paris Principles in 1961 (Lubetzsky 1963). The traditional library collocation is attained by clustering item descriptions by agent names (e.g., authors) and titles – enabling this collocation is a major contribution of the Library cataloger. These clusters are, of course, done by indexing – in the past via the card catalog, but now via machine. Authors' names may vary, work titles may vary, and work content may vary but bringing together descriptions using different criteria gives the end user the ability to find the most useful resources for their needs.

Authority files were developed to support the clustering function and they work well for names (agents), even though much can be debated (and is) about categories of names – persons, corporations, families, conferences, real, imaginary, animals, spirits, etc. They can even be distilled to what is recently called "real world objects". Either character strings (labels) or identifiers can be associated with them so they can serve the purpose of collocation of an agent's corpus and enable end users to find content more easily.

Titles are more difficult as the precise content associated with title strings is problematic to equate. The library profession has tried to apply the names model to titles to achieve collocation of content and has worked to establish unique labels that are associated with all items having the same content. These are the uniform titles of AACR2 (Anglo-American Cataloguing Rules 1978) and earlier cataloging rules and they were entered into name authority records augmented for titles – where additional data included alternate labels (i.e., references) for the uniform title. These title authority records do not contain descriptions of the contents the titles represent, but leave that to the bibliographic records for the resources. They do, however, contain title character strings or identifiers, like name authorities, and enough information to perform the same clustering or collocation functions as names do.

With the development of FRBR (Functional Requirements for Bibliographic Records 1998), however, a very close look was taken at the data in a bibliographic description to sort out data that could be associated with the conceptual work, the expression of the work, the manifestation of the work/expression, and the item. This dissection of description has been valuable to increase understanding of the bibliographic description, even though strict designation of data elements to work, expression, manifestation, and item does not hold up with the variety found among bibliographic resources – different media, editions over time, uniqueness of expression, rareness, etc.

The FRBR work concept and the authority file uniform title need to be reconciled for a future that can employ the new analysis in a useful way. This has led to an attempt to make the title authority record in MARC (MARC 21 Formats 2020) a FRBR work record; and an attempt, initially, to literally follow FRBR (as contained in RDA 2010) in BIBFRAME (BIBFRAME, n.d.)[1]. In both cases adjustment had to be made to enable fluidity between MARC and BIBFRAME.

---

[1] BIBFRAME is a data model and ontology for bibliographic description. It is designed to replace the MARC standards, and to use linked data principles to make bibliographic data more useful both within and outside the library community.

JLIS.it

MARC Title Authorities

The MARC Authority format (MARC 21 2021) was developed (and has been used for over 40 years) to establish and share authoritative labels for names that could be used across a file to enable collocation of resources associated with the name – creative contributions by the named agent or a subject association of the named agent. MARC authorities focused on including alternative forms of the label (MARC 4XX fields). The ideal is/was that every name used in a bibliographic description would be represented by an authority record and that form was to be used in the bibliographic records for access points.

The authority record concept was also extended to titles. The authority records for titles are different and more complex than those for names. Also title authority records are not made for all titles in a file so they share collocation duties with MARC 245 titles on bibliographic records. Title authority records are usually made when references are needed (1-4 below) or the cataloger wants to add cataloger research information (5-6 below). Title authority records are made for the following special situations:

1. When there are likely to be multiple bibliographic resources that are judged to have the same content and different titles.
2. When there are variations in a title authority label. These may be the title in other languages or scripts, or other editions, for example.
3. When there are joint creators or other related agents. The title authority records them as "alternative titles".
4. When catalogers needs to record related titles that have a special association with the authorized title.

In addition, over time notes were added to record:

5. Supporting information for the formulation of the title label.
6. General notes about the title.

At the Library of Congress, title authority records are also generally made for titles for which the Library does not hold the resource but the title is needed in a MARC bibliographic record as an added entry or as a subject. Since the Library of Congress does not have the related resource, there is no bibliographic record for it in the Library of Congress files so the MARC title authority record is a stand-in for a MARC bibliographic record for the related title.

With these "rules" for when a title authority is made, only a small number of title authority records are made. At the Library of Congress while there are over 21 million titles in the bibliographic file, there are only 1.5 million title authority records. It should be noted that title authority records are not made for many cases where a relationship is expressed by a simple added entry. In those cases the bibliographic record serves the authority record role.

Recently attempts have been made in the community to make the MARC title authority serve as a FRBR/RDA work record, which has resulted in proposals to add many elements from the MARC bibliographic format to the MARC authority format to accommodate the additional FRBR work

JLIS.it

elements – effectively making the MARC authority an authority/bibliographic record. This is not easy to do, however, as the tag groups in the MARC authority format are not compatible with those in the MARC bibliographic format.

The Library of Congress undertook as internal study in 2018 to map the MARC title authority record elements used for title authorities to a MARC bibliographic record to see if it was feasible and less disruptive to simply use the MARC bibliographic format for the title authoritative label records. This would have the advantage of enabling libraries to use additional elements for the bibliographic description of a work if an institution wants to add them, rather than using inappropriate fields in the MARC authority format for the data. It would also avoid a massive undertaking to add the missing elements to the MARC authority format. The study found a good fit for the title authorities with only a few adjustments.

The Library of Congress could also see that this would enable a more fluid transformation between formats – with, of course, BIBFRAME being a primary consideration.

BIBFRAME Hubs

When the first pilot for BIBFRAME began at the Library of Congress an attempt was made to use the FRBR/RDA model. BIBFRAME took a slightly simplified approach to FRBR and combined work and expression. The FRBR manifestation was called an "instance" to keep it from being mistaken for equivalence to a FRBR manifestation, although the two were closely aligned. While simplified, the BIBFRAME work/expression and instance shared many of the characteristics of the FRBR/RDA model entities. The Library of Congress began testing this RDF-based ontology with a pilot program, Pilot 1.

Sorting data elements and collecting relationships

For Pilot 1 an attempt was made to identify the data elements in MARC bibliographic records that FRBR/RDA associated with a work/expression and those it associated with an instance. When converting MARC records to BIBFRAME descriptions this allocation of data was made by machine. However, "well curated" as Library of Congress data is it has a long history that includes different sets of cataloging guidelines (ALA, AACR, AACR2, RDA to name a few)[2], community practices, and internal Library of Congress policies that affected consistency across a file of 21 million records. Those records describe resources from text to maps, audio-visuals, music, and still images – in print and various electronic forms. The files of records have been continuously added to for the last half century – with large numbers of records being added from retrospective conversion of catalog cards carried out 40 years ago using minimal record guidelines and then massaged in various projects to improve them.

---

[2] The primary rules used by the Library of Congress since 1908 include: *Catalog Rules: Author and Title Entries*, 1908; *American Library Association rules: A.L.A. Cataloging Rules for Author and Title Entries*, 1949 (ALA); *Library of Congress rules: Rules for Descriptive Cataloging in the Library of Congress*, 1949; *Anglo-American Cataloguing Rules*, 1967 (AACR); *Anglo-American Cataloguing Rules, 2nd ed.*, 1978 (AACR2); *Resource Description and Access*, 2010 (RDA).

JLIS.it

Yet, the BIBFRAME system had to rely heavily on label matching to establish relationships and identify the proper URIs for data found in the MARC record. The system exploited some relationships that originated in the MARC bibliographic linking entries in the MARC 76X-78X, that sometimes have slightly more data to identify links. Many others came from added entries in MARC 700-740. And, of course, the prime MARC links in bibliographic records, the MARC 130 and 240 uniform titles were used. Series entries in the MARC 800-830 produced additional relationships between bibliographic resource descriptions as did 6XX subject entries. These relationships created collocation in the catalog so they were a key focus in the conversion to BIBFRAME. The relationships were collected into "hubs" and it was quickly realized that the hub provided additional power to the BIBFRAME file in support of collocation.
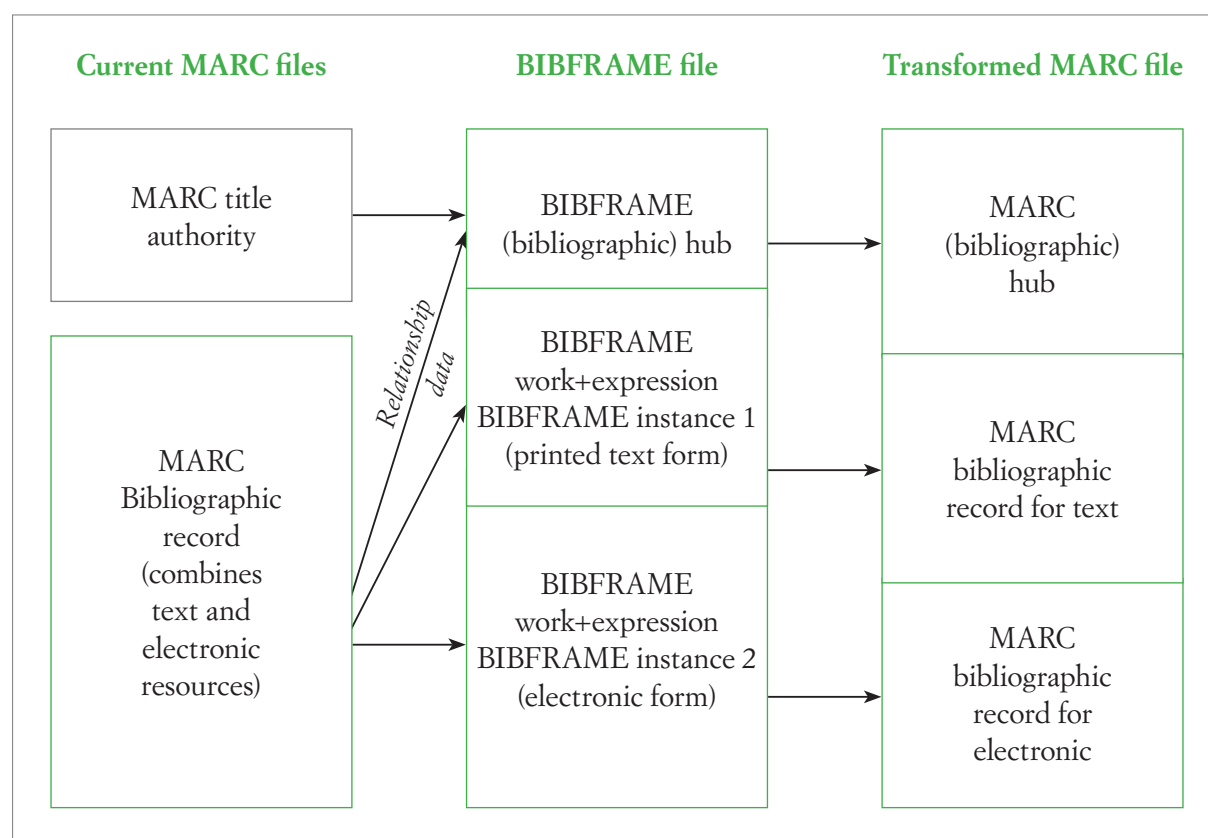
Fig. 1. Current MARC files, BIBFRAME file, transformed MARC file

Despite this exploratory effort creating hubs, Pilot 1 focused on merging, or trying to merge like bibliographic descriptions, or records, when the same resource was described. A difficult aspect of this merging was bringing together subject headings when multiple MARC bibliographic records merged to create one BIBFRAME work description. The subjects were considered part of the work description according to the FRBR/RDA model, not instance properties. Thus, when several MARC records collapsed into one BIBFRAME work, an attempt was made to reconcile the subjects. The merging of subjects proved to be especially difficult.

Pilot 2 and Hubs

So, when the Library of Congress started its second pilot, Pilot 2, it was based on lessons learned in Pilot 1 (BIBFRAME 2016) which included augmentation of the BIBFRAME ontology to better reflect aspects of RDA. But more importantly the project moved to a more realistic model that used "hubs" for collocation based on experience from Pilot 1, allowing the pilot to realize or take advantage of the collocation that had been provided in the MARC environment with the title authority records. The MARC title authorities were converted to BIBFRAME bibliographic work descriptions and called hubs, providing a solid foundation for hubs. Those 1.5 million hubs were then added to when hubs and relationships were created from the MARC bibliographic records as described above, bringing the total to more than 2.3 million. The BIB-FRAME hub is a BIBFRAME bibliographic entity, not an authority description, and our current direction is – starting from the point of view of a BIBFRAME hub – to align the BIBFRAME hub with the MARC bibliographic format, not the MARC authority format as has been library practice. Our work with hubs has clarified a long-standing issue: the title authority is really a bibliographic record in authority clothing! This is a step toward the fluidity needed between BIBFRAME and MARC.

The expanded hub contains data that would have resided in a MARC title authority. It contains the title variations, author/title labels when there are multiple creators, and cataloger notes that support the hub content. And it has some characteristics of a work description. But it will not contain subject information allocated to the FRBR work, which will remain in the BIBFRAME work description, thus avoiding the merger issue. However, the BIBFRAME bibliographic ontology that is used for hubs can easily support further development of the hub description. Because of their similarity to a BIBFRAME work, currently hub descriptions are being expressed as BIB-FRAME works with a special "rdf:type" of hub, which will allow the extension of hub content as needed to include new differentiating elements.

Hubs function as authoritative resources designed to serve as a common denominator, control point, and collocation mechanism, but that is not to say that they are "authorities" and should live separately from the larger bibliographic file. That is what happens now in the MARC files because the format it resides in is the MARC authority format. The format of its storage has dictated how they are seen and where they live. What is being proposed here is not to make these resources any less authoritative and representative than they are today, but to merge them with like data – all bibliographic – to improve their efficacy. The association of the hub with the bibliographic concept is working well thus far in the BIBFRAME environment.

As catalogers can originate more descriptions in BIBFRAME, the hub concept no doubt will continue to develop, but that development will be in a new environment that understands and exploits linking.

JLIS.it

Hubs, SuperWorks, Opuses

There are currently several major projects carrying out extensive implementations of BIBFRAME in an Open Linked Data environment. Several have realized similar needs to those the Library of Congress sought with its hubs. Prominent among the projects is one called Share Virtual Discovery Environment (Share-VDE), a collaborative endeavor of the international bibliographic agency Casalini Libri and @CULT, together with library groups in the United States, Canada and Europe (Share-VDE, n.d.). Share-VDE uses a concept similar to the hub, which they call the "Opus". Another is the University of Alberta's LD4P project[3] where the concept was also given the name "opus". It is meaningful that several projects in the linked data space wrestling with the same problems have developed more or less the same solution.

Going Forward

This paper has discussed some fundamental concepts in bibliographic control in relation to widespread practices in bibliographic description. As the bibliographic environment shifts to take increasing advantage of linked data opportunities, flexibility and fluidity are going to be important. Movement between system environments rooted in MARC and those based in BIBFRAME are essential so narrowing selected differences are important. Discussion will be needed for the community to shift MARC title authorities to MARC bibliographic hubs in synch with BIBFRAME hubs, but in keeping with its commitment to cooperation in the bibliographic world the Library of Congress will pursue that discussion.

---

[3] The University of Alberta is a cohort in the Linked Data for Production (LD4P) project. LD4P is a family of successive grant funded (Mellon) projects that provided foundational work and continued with implementation phases in support of the library cataloging community's shift to linked data for the creation and manipulation of their metadata.

JLIS.it

References

Anglo-American Cataloguing Rules. 1978. 2nd ed. Chicago: ALA.

BIBFRAME. 2016. *"*BIBFRAME Pilot (Phase One—Sept. 8, 2015 - March 31, 2016): Report and Assessment." https://www.loc.gov/bibframe/docs/pdf/bibframe-pilot-phase1-analysis.pdf.

BIBFRAME. n.d. *"*Bibliographic Framework Initiative." Accessed 29 November 2021. https://www.loc.gov/bibframe/.

Cutter, Charles A. 1889. *Rules for a Dictionary Catalogue.* 2nd ed., with corrections and additions. Washington, D.C.: Government Printing Office.

Functional Requirements for Bibliographic Records. Final Report. 1998. Munich: K.G. Saur. https://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

Lubetzky, Seymour. 1963. "The Function of the Main Entry in the Alphabetical Catalogue: One Approach." In *International Conference on Cataloguing Principles, Paris, 9th-18th October, 1961. Report.* London: International Federation of Library Associations. 139-143.

"MARC 21 Formats." 2020. Washington: Library of Congress. Last modified 13 March, 2020. https://www.loc.gov/marc/.

"MARC 21 Format for Authority Data." 2021. Washington: Library of Congress. Last modified 24 November, 2021. https://www.loc.gov/marc/authority.

RDA: resource description and access. 2010 Chicago: American Library Association.

Share-VDE. n.d. "Share-VDE virtual discovery environment." Accessed 29 November 2021. https://share-vde.org/sharevde/info.vm.

JLIS.it

Universal bibliographic control in the semantic web.
Opportunities and challenges
for the reconciliation of bibliographic data models

Tiziana Possemato[a]

a) Università degli Studi di Firenze, http://orcid.org/0000-0002-7184-4070

ABSTRACT

The principles and conceptual models of universal bibliographic control and those of the Semantic web share the common goal of organizing the documentary universe by highlighting relevant entities and mutual relationships, in order to ensure the widest possible access to knowledge. This drives a significant change in the entire information chain, from the analysis and structuring of the data to their dissemination and use. From the construction of bibliographic data models, the point of view, the semantic web paradigm pushes the boundaries of the exchange of records among relatively homogeneous cataloguing systems and opens a transversal dialogue between different actors and systems, in a digital ecosystem that is not contained within cultural, linguistic, geographical or thematic limits. In this context, it is necessary to dialogue with heterogeneous communities of varying authority, driven by the web and often created by institutions or groups of users quite different from the ones to which cataloguing tradition is accustomed. The free reuse of data can also take place in very different contexts from those of their origin, multiplying for everyone the opportunities for universal access and the production of new knowledge. Can different cataloguing traditions coexist in such a changed context and integrate without losing their information value? Based on some recent experiences, this appears to be possible.

JLIS.it

Es ist die Maja, der Schleier des Truges, welcher die Augen der Sterblichen umhüllt und sie eine Welt sehn läßt, von der man weder sagen kann, daß sie sei, noch auch, daß sie nicht sei: denn sie gleicht dem Traume, gleicht dem Sonnenglanz auf dem Sande, welchen der Wanderer von ferne für ein Wasser hält, oder auch dem hingeworfenen Strick, den er für eine Schlange ansieht.
Arthur Schopenhauer, *Die Welt als Wille und Vorstellung*

Background

The 1970's IFLA Universal Bibliographic Control and International MARC (UBCIM) office can be considered the starting point for a larger discussion about Universal Bibliographic Control: it defined some core items, such as the importance of the international sharing of bibliographic data to help reduce costs and to encourage greater cooperation worldwide. The aim was that each national bibliographic agency would catalog the works published in its own country and establish the names of its authors, and that the data would be shared and re-used around the world. Under the theoretical UBC, any document would only be catalogued once in its country of origin, and that record would then be available for the use of any library in the world. In 1974 Dorothy Anderson publishes *Universal Bibliographic Control: a long term policy – A plan for action*, originally prepared as a working document presented by IFLA to the Unesco Intergovernmental Conference on the Planning of National Overall Documentation, Library and Archives Infrastructures, which was held from 23 to 27 September 1974. The document emphasizes the responsibility of national bibliographic agencies to create an authoritative bibliographic record of domestic publications and to make them available to other bibliographic agencies. The process is carried out only by following international standards, in the creation of both bibliographic and authority records (Gordon and Willer 2014). In it, some items are clearly underlined:

1. the responsibility of national bibliographic agencies for creating an authoritative bibliographic record of publications from their own countries;
2. the need to follow international standards in the creation of both bibliographic and authority records.

As Dorothy Anderson affirms "Under the title Universal Bibliographic Control (UBC) IFLA is proposing that Unesco adopts as a major policy objective the promotion of a world-wide system for the control and exchange of bibliographic information. The purpose of the system is to make universally and promptly available, in a form which is internationally acceptable, basic bibliographic data on all publications issued in all countries" (Anderson 1974, 11).

In the foreword of the UBC publication, Herman Liebaers, President of IFLA in 1974, gives an historical background of the context in which the UBC was born: the watershed in the conception of a concretely international approach to collaboration between institutions was given by World War II. Before World War II, institutions expressed international inspirations but were held back by evident technological limitations; project with an international vocation – not only related to librarianship – were proposed and discussed in international context, yet still with a strongly nationalist

JLIS.it

approach and conception. It was only after WWII that the library community, as well as many other professional communities, found itself dealing with the international technological revolution, which completely transformed it. This transformation was absorbed and made its own by IFLA which, from a sort of amateur club of leading European librarians, became "an international professional association prepared to take the lead in policy and in action to serve the library community. It also discovered that at the international level an organization cannot build on national strengths alone, but also to take account of regional weaknesses" (Anderson 1974, 5).

The result of this new IFLA maturity is the UBC program. What is immediately evident was the fact that many concepts concretely expressed in the UBC Program already existed before this formulation. And when the LC announced its Shared Cataloging Program at the IFLA General Council in The Hague in 1966, the impression was that so many of the concepts and issues expressed there already existed in the library community, even if not explicitly expressed. As Herman Liebaers recalls in the same Foreword to Dorothy Anderson's work, Carlos V. Penna, a UNESCO official, after listening to the presentation of the Library of Congress's Shared Cataloging Program exclaims: "but this is universal bibliographic control". The conclusion of Liebaers' same preface is significant in expressing the heart of the UBC as it is now formally defined: "UBC may appear to offer at the technical level of librarianship a balance between humanities and sciences in any new society which is under construction. In its essence UBC is no more than a specific expression of that continuity of knowledge, experience and wisdom for which libraries have always existed" (Anderson 1974, 7).

While the concept of Universal Bibliographic Control was maturing, a crucial moment was constituted by the theoretical and technological ferment that was produced towards the 1990's: the extension of resource formats, with the relative cataloguing rules and standards[1] combined with the centrality of the user's needs brought out the importance of having understandable data "locally", even in a world of shared data. It was recognized that having data in their own languages and scripts, users could understand them; this is extremely important, and by doing so, respecting the cultural diversity of users around the world should be addressed as well. This aspiration was welcomed and accompanied by new web technologies, which however opened the frontiers to another binomial: the relationship between *local* and *global* dimensions and their balance. Web technologies offer new possibilities for sharing data at a global scale and beyond the library domain, but also show a need for *authoritative* and *trusted data*.

In 2008 the Library of Congress Working Group on the Future of Bibliographic Control published the Report *On the record*, that seemed to start from the milestones already defined by Tim Berners-Lee in his linked data design (Berners-Lee 2006)[2]. Some of the most significant themes featured in the report were:

---

[1] Interesting is the evolution from ISBD(CF) to ISBD(ER), to express the urgent exigence to manage electronic resources for the large extension of this kind of resource. See how this evolution is outlined by Stefano Gambari and Mauro Guerrini (Gambari and Guerrini 2002, 75-76).

[2] Tim Berners-Lee outlined four principles of linked data, paraphrased along the following lines:
   1. Uniform Resource Identifiers (URIs) should be used to name and identify individual things.
   2. HTTP URIs should be used to allow these things to be looked up, interpreted, and subsequently "dereferenced".
   3. Useful information about what a name identifies should be provided through open standards such as RDF, SPARQL, etc.
   4. When publishing data on the Web, other things should be referred to using their HTTP URI-based names.

JLIS.it

- the transformation of textual description into a set of data usable for automatic processing by machines;
- the need to make data elements uniquely identifiable within the information context of the web;
- the need for data to be compliant with web technologies and standards;
- the need to use a transversal and interoperable language in the reality of the web.

The *On the record* report officially declares the need to adopt, in the definition of standards and rules, new web technologies and related languages, in order to evolve from a rigid, monolithic language and limitation to the domain (MARC in all its declinations) to something open and comprehensible on a global level (the wider web). This is an important and highly influential reflection for an in-depth re-foundation of Universal Bibliographic Control which, as it encounters new web technologies and the more general paradigm of linked open data, must modify itself to continue to make sense in a web of information that reaches much further than any single, national or international domain of knowledge (Working Group on the Future of Bibliographic Control 2008).

So, assuming that this whole context was the cultural and technological substratum for a new vision of bibliographic control, in December 2012 IFLA reaffirmed the different but closely related positions and roles of IFLA and National Bibliographic Agencies (NBA) in the context of Universal Bibliographic Control. IFLA's vision was expressed through the following principles:

- A National bibliographic agency (NBA) has the responsibility for providing the authoritative bibliographic data for publications of its own country and for making that data available to other NBAs, libraries, and other communities [...]
- NBAs, as a part of the creation of authoritative bibliographic data, also have the responsibility for documenting authorized access points for persons, families, corporate bodies, names of places, and authoritative citations for works related to its own countries [...]
- IFLA has [...] the responsibility for creating, maintaining and promoting bibliographic standards and guidelines to facilitate this sharing of bibliographic and authority data (e.g., ISBD, the FRBR family of conceptual models, etc.);
- IFLA works collaboratively with other international organizations (e.g., ISO, ICA, ICOM, etc.) in the creation and maintenance of other standards in order to ensure that library standards developments, including compatible data models, are coordinated with those of the wider community.[3]

Think global, act local

The National Bibliographic Agencies thus approach their fundamental role by pursuing a number of important issues and paying particular attention to specific themes, including:

- production that expresses the cultural richness of one's country, be it produced locally or from another country;

---

[3] <https://www.ifla.org/files/assets/bibliography/Documents/ifla-professional-statement-on-ubc-en.pdf>.

JLIS.it

- extension to global content of interest to its users, related (or not) to local content;
- attention to the way the content is expressed through metadata with the application of international standards and rules but with frequent "local" choices (example: the rule of presenting as a favoured the form of a name understandable to your users);
- universal standards and rules applied locally, for specific needs.

The focus is on the NBA's responsibility to provide authoritative bibliographic data for their country's publications and share them with a wider community. The role of the National Bibliographic Agency is to express the cultural richness of a country in a way that can be shared with other countries and agencies, coordinated by IFLA in providing standards and guidelines to make data universally shareable, in a global community. The two-dimensional vision of local production in a global context is evident: the popular remark made by Patrick Geddes "*Think global, act local*", probably used originally in city planning and extended in many wider contexts including the environment and culture, seems to match exactly with the new aspiration expressed by IFLA and NBAs.

Patrick Geddes' statement seems to definitively express this duality, in cataloguing, between local expression and global aspiration, between local vision and global perspective, which does not only concern the content of what is conveyed by the NBAs, but also the form and therefore the way of expressing them. As Gordon Dunsire and Mirna Willer affirm in their article *The local in the global: universal bibliographic control from the bottom up* "Local content is held in global carriers, and global content is held in local carriers" (Gordon and Willer 2014).

This balance of local and global vision within UBC worked well until the content being broadcast was defined by National Bibliographic Agencies and controlled through descriptions (metadata), built in compliance with shared rules and standards. All expressed through bibliographic and authority records. The *record* maintains its position as absolute protagonist and conveys this dual trend quite effectively. The Marc format, which can be declined into various dialects of the same family, has largely contributed to creating an object around which services have been built and has, at times, become something that can condition cataloguing choices even more than the rules themselves, giving rise to the expression "cataloguing in Marc" instead of cataloguing according to one of the existent cataloguing rules and guidelines. From an *exchange format* it has become a *cataloguing format* to the point that in many public calls for the acquisition of cataloguing software the constraint "cataloguing must take place in Marc" is, in a technically misinformed sense, usually included. This enormous success is also evidenced by its long duration and the investments made to keep it constantly updated in order to keep pace with the requirements of users and institutions, while always showing some difficulty in getting out of the domain of librarianship.

From identity to entity: the veil of Mâyâ

A good story doesn't necessarily last forever: the record, after almost 60 years of widespread use within the library community, has begun to show its limits in comparison with the languages of the web, which are lighter, partially more understandable and above all transversal (Tennant 2002, 26-28). The record, both bibliographic and of authority, is traditionally rich in information,

JLIS.it

readable by machines but still not "understandable" to them: it maintains the characteristic of being a flat, auto consistent[4] description of an object but not the object itself, not the Real World Object (RWO) that has taken the leading role in the new dimension of the semantic web (Coyle 2015). So, in the context of cataloguing approaches, the record becomes again a protagonist of a new revolution: from the *record*, as a whole with meaning in its entirety, to *entities* as real things in the world, as Real World Object. Each record has metadata that are useful to derive properties in order to build entities. But they are hidden and usually expressed in a way that only partially represents the entity, which could be expressed in various ways.

The language of the web runs in support of traditional standards in order to simplify the information and make it understandable. The goal is to have a method so simple that it can express anything and at the same time so structured that it can be used – and reused – by computer applications: the Resource Description Framework model,[5] in its extreme simplicity of a triple (a subject – a predicate – an object), able to express everything, seems to respond to the need to make data globally shareable, understandable, reusable, in a wider and cross-domain environment. This new perspective is not reducible only to a change of format or technologies, but it expresses a change of approach in the vision of the world: it is a new, umpteenth attempt by humanity to bring the heart of things closer, to go beyond mere representation of them and get to grasp their essence. But the description of things, despite all attempts to go beyond appearance, means giving a *representation of reality*. The new languages of the web express the attempt to bring down the veil of Maya, the one that obscures the sight of humans and does not allow them to reach reality:

> It is Mâyâ, the veil of deception, which blinds the eyes of mortals, and makes them behold a world of which they cannot say either that it is or that it is not: for it is like a dream; it is like the sunshine on the sand which the traveller takes from afar for water, or the stray piece of rope he mistakes for a snake.

This epochal transition from strings to things, from a description to an entity, was largely favoured by the linked open data paradigm and by the new way of understanding and structuring data, decisively shifting the focus from identity, as a form of presentation of an entity, to a real entity, consisting of a series of properties and relationships useful for its identification. The long cataloguing tradition, with its rules and standards that have followed one another over time and that have guided the cataloguing choices, both semantic and syntactic, was born and raised on a distinction between entity and identity (one entity, many identities) that was never clearly defined. Although seen as a simplification, the definition of identity (as a philosophical concept) in its rela-

---

[4] The Marc record, with its Directory that clearly expresses it as a whole, has a meaning and a value in its entirety: each element of the description, outside the record itself, loses meaning and identity.

[5] "RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This graph view is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations." <https://www.w3.org/RDF/>

tionship with an entity, proposed by Wikidata, is meaningful: *"Identity is all that makes an entity definable and recognizable, because it possesses a set of qualities or characteristics that make it what it is and, for that very reason, distinguish it from all other entities".* [6] This transition in the cataloguing approach can be seen as a shifting from identity, as a form of presentation of an entity, to a real entity, consisting of a series of properties and relationships useful for its identification.

The cataloguing tradition has for centuries been focused on the record intended as a synthesis of the expression of an identity. Behind the topos *"Are the winner of Austerlitz and the loser of Waterloo the same person?"* there is the meaning of this philosophical but also practical passage: behind the many possible expressions of an identity there is a unique and, in some ways, unrepeatable entity.

"Are the winner of Austerlitz and the loser of Waterloo the same person?"

Fig. 1. The entity Napoleon is represented by many identities

The world is my representation

The shift of attention from the record to the entity, understood as a Real World Object, could be represented as the passage from a flat, static, 2-dimensional worldview to a dynamic, 3-dimensional worldview. In cataloguing terms, we are facing a crucial transition from a representation of the world, to the world in itself, in its concreteness and variety, and to the attempt, which remains so, to express it in its reality. However faithful or authoritative the description is, it always remains a *representation* of a reality, which is other than reality itself. But the change of view helps the observer to get closer to that reality and to interpret it in a different way, hopefully, more respectful of the object represented: this is easily and visually expressed as the passage from a flat, static, 2-dimensional worldview to a dynamic, 3-dimensional worldview. The record, often expressed through a globally shared syntax, but within specific communities and specific domains, manifests all the limits of a monolithic and flat object: the resource told through the traditional bibliographic or authority record, is as if it assumed the same two-dimensional and static features of its representation.

---

[6] <https://it.wikipedia.org/wiki/Identit%C3%A0_(filosofia)>

JLIS.it

Fig. 2. Van Gogh's portrait with its different descriptions in Marc records

The transition to the Real World Object refers to another way of understanding the object, in its three-dimensionality and concreteness. Those who produce metadata are still obliged to remain on this side of the veil of Mâyâ that we were talking about, but they come close to a three-dimensional object, which can be observed from a variety of points of view. A view that best allows you to tell the "thing" (Thing) in its being a thing (a work, a person, a place, an abstract concept...). The view of the producer of the metadata becomes the same view of whoever (in the example of Van Gogh's portrait) tries to look at it as the original creator must have seen and imagined it, thus approaching what his idea should have been originally, although still being able to give "only" one (or more) representations.

Entity is built by putting together properties expressed through different ontologies and vocabularies, from different institutions. And the same entity, the Real World Object, can continue to be expressed through one or multiple identities.

JLIS.it

Fig. 3. Van Gogh's 3D view, with properties expressed through different ontologies and vocabularies

What is striking in this new perspective is the change in the cataloguing geography: new sources, not necessarily institutional, can contribute to represent the same entity, in a network that goes beyond local or national borders creating a digital ecosystem that is not at all contained within cultural, linguistic, geographical or thematic limits. We are living on a cloud of data: many domains meet on the web to enrich and extend the informative power of data. Libraries are, in a certain way, forced to reorganize themselves in a similar way, proposing a wider network where each node can be constituted by a library, an archive, a museum or any information provider. In this context, it is necessary to dialogue with heterogeneous communities of varying authority, driven by the web and often created by institutions or groups of users quite different from the ones to which the cataloguing tradition is accustomed. The purpose of this cooperation between different domains is articulated: it includes the possibility of making data creation and management processes sustainable in the long term with the ability to enrich data using different sources and reuse something that was not originally created in its own domain, without any political, cultural and technological barrier. The free reuse of data can take place in very different contexts from those of the origin of that data, multiplying for everyone the opportunities for universal access and the production of new knowledge.

To give a clearer idea of the wealth of standards and metadata limited to the cultural heritage sector alone, which can be used to build and format data, ten years ago Jenn Riley published a metadata map: it provides an impressive representation of the standards for the digital collection (105 standards) (Riley 2009-2010). We can only imagine how this map and its relationships can expand out of the limited cultural context and meet with the standards and languages of other communities. In such a broad, complex and heterogeneous ecosystem, which is not always authoritative, does UBC still make sense and do the national agencies that take charge of it still have a role?

JLIS.it

Can different cataloguing traditions coexist in such a flowing context and integrate without losing their information value and authoritative character?

Anyone can say Anything about Anything

Each ontology or dataset refers to an institution or a community, with its strength and authority guaranteed, for the most part, by the strength and authority of the community responsible for creating and managing this source. The strength of a community, which guarantees the *authoritativeness* and *certifiability* of a source, is also given by the number (*quantitative* aspect) and by the typology (*qualitative* aspect) of the community guarantor of the source. These precepts should partially stem the risks inherent to the AAA Principle, which is the founding base of the Semantic Web: *Anyone can say Anything about Anything.*[7]

But if it may seem rather simple to frame, verify and certify the quantitative data of a community that supports and produces a source, through measurement criteria, the evaluation of the qualitative data is not so easy. And this is all the truer if we think of a global dimension, such as that of the web, in which a community can be spread beyond any possible measurable boundary. It is here that the concept of authority risks having to give way to the concept of *consensus*, and it is here that, perhaps, even more so, we need to rethink and strengthen the concept of certified authority of a source.

As Giovanni Pirrotta writes, data constitute the skeleton upon which the structure of communication is built. The more the data is authentic, truthful, authoritative, certified and verifiable, the more difficult it is to invent fake news (Pirrotta 2019). In his article, Pirrotta tries to demonstrate that it is possible to certify and verify data also with the support of new web technology. Using authoritative sources, he demonstrates that the possibility for a machine to cross different sources and to certify data is possible and it constitutes a way of getting proof and giving trust to an assertion.

In the example of figure 4, the entity *Elio Morpurgo*, an Italian politician of Jewish origin, a victim of the Holocaust, is rebuilt through highly authoritative sources:

- CDEC - Ontology of the Fondazione Centro di Documentazione Ebraica Contemporanea[8]
- OCD - Ontology of the Italian Camera dei Deputati[9]
- Ontology of the Italian Senato della Repubblica[10]

The sources used to identify the entity are created and maintained by very authoritative institutions, able to assure the quality and the accuracy of the data: the truthfulness of the information depends on the quality of the source.

---

[7] "To facilitate operation at Internet scale, RDF is an open-world framework that allows anyone to make statements about any resource. In general, it is not assumed that complete information about any resource is available. RDF does not prevent anyone from making assertions that are nonsensical or inconsistent with other statements, or the world as people see it. Designers of applications that use RDF should be aware of this and may design their applications to tolerate incomplete or inconsistent sources of information". <https://www.w3.org/TR/rdf-concepts/#section-anyone>

[8] <http://dati.cdec.it/>

[9] <http://dati.camera.it/ocd/reference_document/>

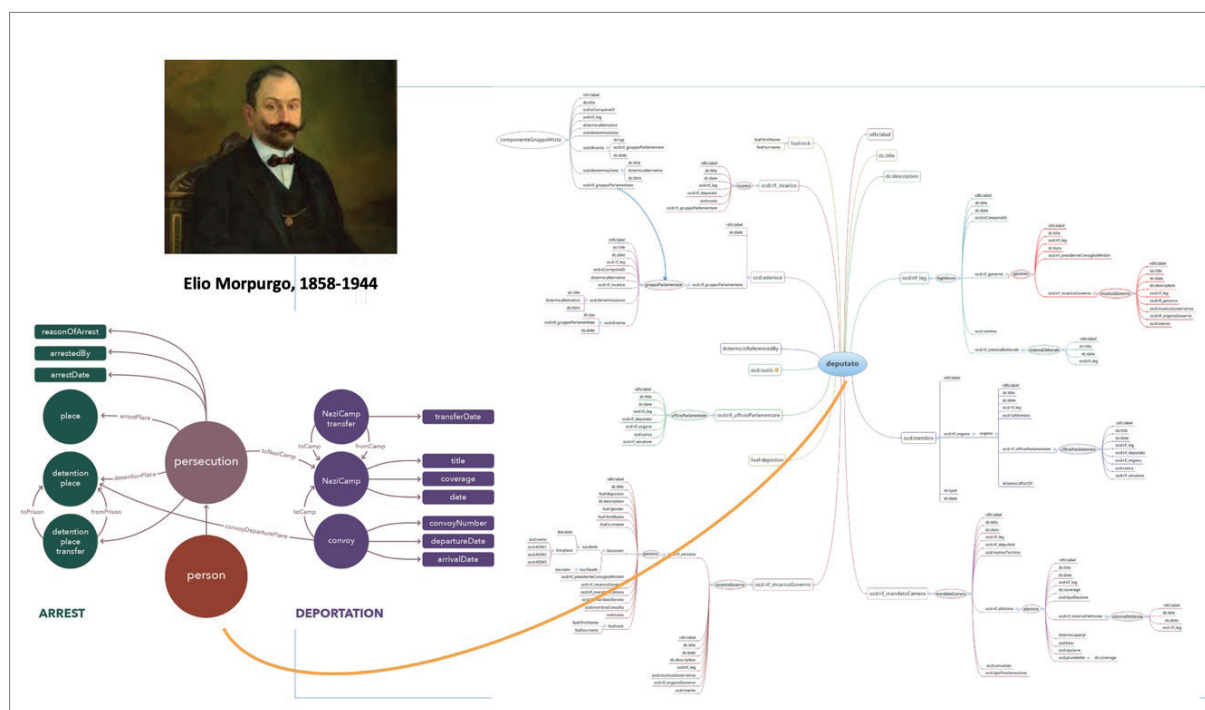[10] <http://dati.senato.it/sito/21>

JLIS.it

Fig. 4. The Italian politician Elio Morpurgo is identified through highly authoritative sources

In such a complex information production chain, where data are built and distributed by heterogeneous sources, ranging from authors to publishers, from blogs to libraries, from social networks to ontologies, the role of National Bibliographic Agencies cannot fail to become central; their contribution in terms of authoritativeness remains fundamental, and indeed, acquires centrality again in a new global scenario in which each source can contribute to building the most effective representation of an entity, but many sources cannot guarantee the character of *authority*, *persistence* and *updating*. In this transition phase, where processes are evolving from bibliographic and authority control to entity management in a shared environment, where it seems that the strongest approach is the AAA Principle, where authority seems to have been superseded by consensus, a founding criterion has to be defined to harmonize the different voices speaking about a thing: *democracy* seems the most effective way of balancing and coordinating a community in which anyone can say anything about anything.

How this principle is applied to building entities and how it affects entity identification and presentation strategies can be briefly summarized as follows:

- each National Bibliography Agency can choose any preferred form and all variants to identify or to present an entity (it can choose the number and type of its attributes); the constraints on the formats lapse;
- all "locally preferred" forms have become equal in a globally shared environment, in a cluster of variant forms that is not affected by hierarchical structure and logic.

But, as in all democratic systems, it is necessary to choose someone who represents people; thus, even in the representation of entities, different institutions can choose from among different vari-

ant (literal) forms the one that best represents the entity in their own community, in order to better meet the needs of its users (whether cultural, geographical, domain, linguistic needs, etc.).

Fig. 5. The form of the name for Cicero, chosen by the Biblioteca nazionale centrale di Firenze and by the National Library of Estonia

As clearly expressed in the AAA Principle, the RDF model used to structure data in the semantic web does not presuppose and guarantee that the assertion is correct in the message conveyed, but that it is formally well structured, with a subject, a predicate, an object. RDF does not warrant that nonsense or inconsistent statements will not be made with other statements. Consequently, we are aware that an enormous number of triples are created in the Semantic Web, regardless of their quality and truth.

So, if the assertion expressed by the triple is:
"the Earth – is – flat"
or if the assertion expressed by the triple is:
"the Earth – is – round"
in term of RDF is exactly the same: both are well structured assertions.

In the same way, if the assertion is:
"The preferred label – is – Pirandello, Luigi, 1867-1936"
or
"The preferred label – is – ולדנריפ, י'גיאול, 1867-1936"
it's absolutely neutral for RDF.

The certification of "who says something" is expressed through the *fourth element* – the Provenance – added to the original triple.
Its role, in a shared environment, is fundamental:
- it ensures that each institution, as a source, assumes responsibility for the data (data trust);

JLIS.it

- it allows institutions to share their data in wider contexts, keeping track of their contributions (data traceability);
- it allows users (professionals or end-users, as well as machines) to apply filters to select data from specific sources (application profile).

So, to go back to the example used above, triples become quadruples and declare the responsibility of whoever makes an assertion:

"The ICCU says that – the preferred label – is – Pirandello, Luigi, 1867-1936"
or
"The National Library of Israel says that – the preferred label – is – ‏ילדנריפ, ‏י'גיאול‎, 1867-1936"

In this way, anyone can say anything about anything, assuming the responsibility of the assertion.

Conclusion

The attention of the entire data production chain, from the publisher to the cataloguing and distribution agencies, returns to focus on the real and essential information power of the data, which is structured so as to be universally understood and shared. In this new ecosystem, in this new geography with completely open borders, in which the actors and information elements are themselves open and heterogeneous, the constraints and rigidities expressed in the past by formats, standards, rules of national cataloguing, often closely linked to specific domains, completely lose their meaning. Authoritative institutions, both local and global, reaffirm their role and their centrality, provided they are able to adapt themselves and their services to the runaway evolution of the times. In the allegory of Plato's Cave, people who have lived chained to a blank wall of a cave all their lives, watch shadows projected on the wall from real objects and give names to these shadows. The shadows are the prisoners' reality, but are not accurate representations of the real world. The librarian, like any institution that provides data, should become like the philosopher who is freed from the cave and comes to understand that the shadows on the wall are actually not reality at all. Anyone can try to get to the real world knowing that it will probably remain an attempt, and cataloguing and data providing will remain a description of it. But as accurate as possible.

JLIS.it

References

(Last consultation of the websites: 22th April 2021)

Anderson, Dorothy. 1974. *Universal Bibliographic Control: a long term policy, a plan for action*. Pullach/München: Verlag Dokumentation.

Berners-Lee, Tim. 2006. "Linked Data". Design issues for the World Wide Web." W3C. https://www.w3.org/DesignIssues/Overview.html.

Coyle, Karen. 2015. "Coyle's InFormation: Real World Objects." http://kcoyle.blogspot.com/2015/01/real-world-objects.html.

Dunsire, Gordon, and Mirna Willer. 2014. "The local in the global: universal bibliographic control from the bottom up." http://library.ifla.org/817/1/086-dunsire-en.pdf.

Gambari, Stefano, and Mauro Guerrini. 2002. *Definire e catalogare le risorse elettroniche*. Milano: Editrice Bibliografica.

Gonzales, Brighid M. 2014. "Linking Libraries to the Web: Linked Data and the Future of the Bibliographic Record." *Information Technology and Libraries* 33 (4):10-22. https://doi.org/10.6017/ital.v33i4.5631.

Pirrotta, Giovanni. 2019. "Generazione e verifica di notizie di qualità attraverso il Web Semantico: la storia di Liliana Segre." https://medium.com/@gpirrotta/generazione-e-verifica-di-notizie-di-qualità-attraverso-il-web-semantico-la-storia-di-liliana-6cd81f05e9fe

Riley, Jenn. 2009-2010. "Seeing Standards: A Visualization of the Metadata Universe." http://jennriley.com/metadatamap/.

Schreur, Philip. 2018. "The Evolution of BIBFRAME: from MARC Surrogate to Web Conformant Data Model." 13-07-2018. http://library.ifla.org/2202/1/141-schreur-en.pdf .

Schreur, Philip E., and Amy J. Carlson. 2020. "Bridging the Worlds of MARC and Linked Data: Transition, Transformation, Accountability." *The Serials Librarian* 78:1-4, 48-56. DOI: 10.1080/0361526X.2020.1716584

Tennant, Roy. 2002. "MARC must die." *Library Journal* 127 (17):26–28. http://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/#_ .

Working Group on the Future of Bibliographic Control. 2008. "On the Record: report of the Library of Congress Working Group on the Future of Bibliographic Control". https://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf.

JLIS.it

Control or Chaos: Embracing Change and Harnessing Innovation in an Ecosystem of Shared Bibliographic Data

Ian Bigelow[a], Abigail Sparling[b]

a) University of Alberta Library, http://orcid.org/0000-0003-2474-7929
b) University of Alberta Library, http://orcid.org/0000-0002-2635-8348

ABSTRACT

With the transition from MARC to linked data, how we create and manage bibliographic data is drastically changing. This shift provides increased opportunity to test resource description theory and develop best practices. However, efforts to simultaneously define models for creating native linked data descriptions and crosswalk these models with MARC have resulted in ontological differences between implementers and unique extensions. From the outside looking in this progress may look more like bibliographic chaos than control. This apparent chaos, and the associated experimentation is important for communities to chart a path forward, but also points to a challenge ahead. Ultimately this disparate community innovation must be harnessed and consolidated so that open standards development supports the interoperability of library data. This paper will focus on modelling differences between RDA and BIBFRAME, recent attempts at MARC to BIBFRAME conversion, and work on BIBFRAME application profiles, in an attempt to define shared purpose and common ground in the manifestation of real world data. Emphasis will be placed on the balance between core standards (RDA, MARC, BIBFRAME) and community based extensions and practice (LC, PCC, LD4P, Share-VDE), and the need for a feedback loop from one to the other.

KEYWORDS

BIBFRAME; Bibliographic control; Cataloguing; Linked data; Metadata; Resource Description and Access.

JLIS.it

Introduction: What we will cover

This paper follows closely from the proceedings of the matching presentation at the *International Conference on Bibliographic Control in the Digital Ecosystem* (Bigelow and Sparling 2021). Our goal is to share findings from research and work towards implementation of BIBFRAME, with a particular focus on data exchange and interoperability. Findings are presented with the hope of informing next steps for the cataloguing and metadata standards communities to move forward with core standards supporting bibliographic control in emergent metadata ecosystems.

In an effort to capture some of the challenges for bibliographic control emerging in the changing landscape for library bibliographic metadata we will focus on several key areas of discussion as they relate to data reuse: the intersection of RDA and BIBFRAME; the complexities of historical MARC data through conversion; what standard BIBFRAME and BIBFRAME infrastructure should look like; and in this context how we can harness innovation and maintain control.

Context: Our lens

In 2018 strategic planning at the University of Alberta Library (UAL) resulted in a plan for *Moving Forward with Linked Data* which stated that "In order to reap the benefits of full participation in the linked open data environment, UAL should continue to take steps towards complete conversion of existing library data to linked open data" (Farnel et al. 2018, 8). Since the plan's publication, UAL has continued as a member of the Share Virtual Discovery Environment (Share-VDE) and actively engaged in the Linked Data for Production Phase 2 (LD4P2) as a cohort library. We are also a member of the Program for Cooperative Cataloging (PCC). Much of this paper is informed by experiences and observations as a member of these projects and initiatives.

As such, it is worth noting from the outset that this paper will focus on bibliographic control in a BIBFRAME context. This is in line with decisions at the UAL for transitioning our MARC data to a linked data ecosystem, but also in line with our commitment to the PCC. We fully recognize, however, that PCC does not represent all libraries and that BIBFRAME is just a piece of a larger linked data framework. While much of what we will discuss may have applications for interchange of linked data for libraries as a whole, we have purposely scoped the discussion to BIBFRAME.

Experimentation to Implementation

Leading up to 2018, analysis of conversion from MARC to BIBFRAME was undertaken at UAL (Bigelow et al. 2018). This analysis highlighted that conversion processes captured RDA core elements and were generally functional. Issues were noted however, many of which related to accounting for changes in cataloguing standards over time, and in choices made for mapping MARC to BIBFRAME. We ended the article with a note that "Waiting until we have no choice to transition will not foster the desired community collaboration around BIBFRAME development or support a smooth implementation" (15).

Since 2018, UAL has changed its focus from research and analysis to working towards BIBFRAME implementation. Through work with the LD4P2 Cohort, PCC, and Share-VDE, significant effort

has been put into staff training as well as further refinement of conversion processes, data modelling, and application profiles. BIBFRAME implementation is a large-scale ongoing process that requires revision of our workflows and technical ecosystems to support a hybrid MARC and BIBFRAME environment. As we have undergone this work the importance of replacing workflows for metadata reuse has become top of mind.

Developing workflows for sharing BIBFRAME data presents certain challenges. Testing metadata reuse requires both supporting systems and data sets to share. Now, however, along with the Library of Congress (LC) there are other national libraries (Axelsson 2018; Lendvay 2020) working on BIBFRAME implementation, and numerous other libraries contributing to projects like LD4P (Stanford Libraries 2018) and Share-VDE (Lionetti 2021) such that there are billions of quads of data live in BIBFRAME (Share-VDE 2019). As we know, "Universal Bibliographic Control is grounded on sharing the effort of resource description, eliminating redundancy by encouraging sharing and re-use of bibliographic data" (IFLA 2017). We need to make sure that BIBFRAME data can support interchange. To achieve bibliographic control there needs to be agreement on what standard BIBFRAME looks like.

Interchange with MARC certainly is not perfect. Different communities of practice apply different standards and different MARC formats, quality varies, and the copying of records to local silos duplicates effort. At the same time, systems and practices for working with MARC are so long established that we often take interchange for granted.

Bringing it all together

Beyond the challenges of working with new standards in a linked data environment, the scale of change away from MARC necessitates fairly long term hybrid environments with compounding complexity. Figure 1 is provided as an example, capturing the plans at UAL for linked data implementation.
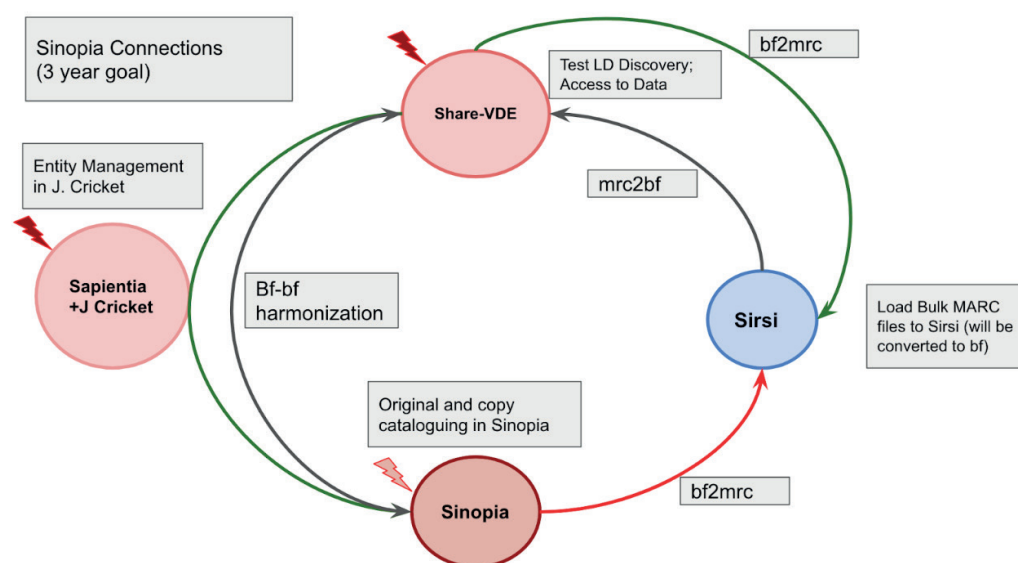

Fig. 1. Sinopia Connections (3 year goal) (Image by Bigelow, 2020)

JLIS.it

While some library systems are beginning to adapt for BIBFRAME, the complexity highlighted in Figure 1 is obvious. Making this kind of transition involves significant adaptation and/or system migration. The scale of such a transition means that not all libraries will be moving from MARC to BIBFRAME at once, necessitating support for hybrid systems for some time. In the case of UAL, current use of SirsiDynix Symphony means that for a number of library services we will still need MARC until a more complete transition is achieved. Moreover, even when we are able to fully transition our own systems, we need to consider the reliance of libraries generally on shared bibliographic data.

As outlined in Figure 1, to work in BIBFRAME we need a cataloguing editor with standardized application profiles with comprehensive coverage to describe a range of resources in BIBFRAME, but we also need conversion and data flow processes established for converting from MARC to BIBFRAME and from BIBFRAME to MARC. One might easily wonder where the problem lies here. After all, multiple MARC to BIBFRAME conversion processes have been established (LC, Share-VDE, LibrisXL, ExLibris), we have the LC BIBFRAME to MARC converter, and both the LC and Sinopia BIBFRAME cataloguing editors. That the library community is now at a point where we have working tools to start putting together a BIBFRAME ecosystem like this is an incredible achievement. On the other hand though, to bridge from individual and project-specific toolings to a functional ecosystem means that they all need to work together, and, given the reliance on shared data in libraries, they don't just need to work together for one institution, but internationally.

With the shift away from MARC for bibliographic description, for the purpose of interchange we are left with two relatively new standards (RDA and BIBFRAME). The combination of these standards is emergent and adds additional complexity to ensuring bibliographic control in a BIBFRAME environment. For the remainder of this paper we will focus on RDA, BIBFRAME and related aspects pertinent to bibliographic control by examining our experiences with LD4P2 and Share-VDE.

RDA and BIBFRAME: Chaos and convergence

To begin wading through the chaotic divide between RDA and BIBFRAME we need to take a trip into the past and the initial release for both standards.

From the very outset of RDA in 2010 there was agreement that an alternative to MARC was required to support the extent of RDA (Cronin 2011; McGrath 2011; Samples 2011). Though MARC has continued to evolve since then, we have now had 10 years where the theoretical underpinnings of RDA have been largely untested by practice. Despite the predominant stasis in encoding standard, RDA has continued to evolve to the point that we have an entirely new version of RDA as of December 2020 (RDA Steering Committee 2020).

BIBFRAME has also had a long development trajectory, beginning in 2011 with the goal of creating a community standard to allow RDA to move beyond MARC. We would argue however, that work on BIBFRAME didn't accelerate with the wider library community until 2017 when LC released conversion tools and specifications for testing. Along the same approximate timeline, early implementation cases for BIBFRAME emerged (Library of Congress, n.d.a), and large scale proj-

ects like LD4P and Share-VDE meant that data and tools in production allowed for development of best practices and testing of theories dating back to when FRBR was initially released in 1998 (Samples and Bigelow 2020; IFLA Study Group on the Functional Requirements for Bibliographic Records, and Standing Committee of the IFLA Section on Cataloguing 1998).

Reflecting on this timeline, 2017-2020 saw increased development not just in BIBFRAME, but in the evaluation, testing and analysis of use of RDA in a linked data environment. This acceleration has resulted in beautiful chaos, with further work on data modelling, more maturity in conversion processes, and use case development driving novel extensions and adaptations. There are a number of excellent articles analyzing how well BIBFRAME can accommodate RDA and associated challenges (Zapounidou, Sfakakis, and Papatheodorou 2019; Taniguchi 2017; Baker, Coyle, and Petiya 2014; Guerrini and Possemato 2016; Seikel and Steele 2020; Taniguchi 2018; El-Sherbini 2018; Zapounidou 2020), and while this is an important question, it is not the only one. With the relative maturity of both standards, and the ability to work with data in live systems, both can now be tested and adjusted to best meet user needs. The question becomes, what does an application profile utilizing RDA and BIBFRAME look like in the real world, and how does it and the data model evolve under the scrutiny of use for resource description and from user feedback?

With the RDA 3R project and the new toolkit, changes to RDA are significant enough that the PCC chose to postpone implementation until at least July 2022 (Program for Cooperative Cataloging Policy Committee 2020). In part this was based on the need for further work on policy statements and metadata documentation, but there was also a recognition that a test is warranted for both application in MARC and BIBFRAME (Ibid.). In 2010 a test was carried out on the application of RDA in MARC, so with the development of BIBFRAME we are only now getting to a point where these many components can come together. As noted in *Exploring Methods for Linked Data Model Evaluation in Practice*, "A final identified way of assessing an ontology involves testing the data itself throughout the modeling process. This could take the form of checking against use cases and competency questions, and user testing of the data in the application" (Desmeules, Turp, and Senior 2020, 68). With implementation cases such as the National Library of Sweden and projects like Share-VDE and LD4P this kind of assessment can finally happen for both BIBFRAME and the use of RDA as a cataloguing content standard with it.

Analysing native BIBFRAME and the use of RDA

Working on the creation of application profiles for the Sinopia cataloguing editor has provided an excellent opportunity to test the application of RDA in BIBFRAME. For this analysis in Sinopia it is worth providing the context that UAL, along with all members of LD4P2 were PCC institutions. While LC application profiles were used as a starting point, Sinopia development then allowed for the creation of base application profiles for all users, and experimentation/localization such that each member could create application profiles of their own. This flexibility continues to be a strength, allowing for ongoing development of core/base application profiles while allowing for testing of new concepts.

JLIS.it

Through the course of work on UAL Sinopia application profiles, decisions on the use of properties needed to be made. In constructing application profiles, thought was given to PCC standards and ensuring that core elements were captured for resource description. While the Sinopia application profiles used for analysis here are UAL specific, they were created in collaboration with LD4P2, the Profiles Affinity Group and with a thought to ongoing work with PCC. The example shown in Figure 2 is an extract of the JSON from the UAL Monographs profile in Sinopia, adjusted into a spreadsheet. Figure 2 presents the property list and labels, the corresponding RDA instruction/entry note, while also reflecting recent modelling updates from Share-VDE.

Resource Template Label	Type	Mandatory	Repeatable	PropertyURI	PropertyLabel	RDA Instruction/Entry Note
UAL Monograph Work (Un-Nested)	resource	false	true	http://id.loc.gov/ontologies/bibframe/expressionOf	Has Opus	
	resource	false	true	http://id.loc.gov/ontologies/bibframe/hasInstance	Has Instance	
	lookup	false	true	http://id.loc.gov/ontologies/bibframe/identifiedBy	Work Identifier	Used with Unspecified
	resource	false	true	http://id.loc.gov/ontologies/bibframe/contribution	Contribution (Creator/Contributor)	
	resource	true	true	http://id.loc.gov/ontologies/bibframe/title	Title Information	
	lookup	false	true	http://id.loc.gov/ontologies/bibframe/genreForm	Form of Work	http://access.rdatoolkit.org/6.3.html
	literal	false	true	http://id.loc.gov/ontologies/bibframe/originDate	Date of Work	http://access.rdatoolkit.org/6.4.html
	lookup	false	true	http://id.loc.gov/ontologies/bibframe/originPlace	Place of Origin of the Work	http://access.rdatoolkit.org/6.5.html
	lookup	false	true	http://id.loc.gov/ontologies/bibframe/geographicCoverage	(Geographic) Coverage of the Content	http://access.rdatoolkit.org/7.3.html
	literal	false	true	http://id.loc.gov/ontologies/bibframe/temporalCoverage	(Time) Coverage of the Content	http://access.rdatoolkit.org/7.3.html
	lookup	false	true	http://id.loc.gov/ontologies/bibframe/intendedAudience	Intended Audience	access.rdatoolkit.org/7.7.html
	lookup	false	true	http://id.loc.gov/ontologies/bibframe/hasSeries	In Series	URI for series as a work
	resource	false	true	http://id.loc.gov/ontologies/bibframe/note	Notes about the Work	
	resource	false	true	http://id.loc.gov/ontologies/bibframe/dissertation	Dissertation	http://access.rdatoolkit.org/7.9.html
	resource	false	true	http://id.loc.gov/ontologies/bibframe/tableOfContents	Contents	
	resource	false	true	http://id.loc.gov/ontologies/bibframe/summary	Summary	http://access.rdatoolkit.org/7.10.html
	lookup	false	true	http://id.loc.gov/ontologies/bibframe/subject	Subject of the Work	http://access.rdatoolkit.org/rdachp23_rda23-12.html
	resource	false	true	http://id.loc.gov/ontologies/bibframe/classification	Classification numbers	
	lookup	true	true	http://id.loc.gov/ontologies/bibframe/content	Content Type	http://access.rdatoolkit.org/6.9.html
	resource	false	true	http://id.loc.gov/ontologies/bibframe/language	Language of Expression	http://access.rdatoolkit.org/6.11.html
	resource	false	true	http://id.loc.gov/ontologies/bibframe/notation	Script	http://access.rdatoolkit.org/7.13.2.html
	lookup	false	true	http://id.loc.gov/ontologies/bibframe/illustrativeContent	Illustrative Content	http://access.rdatoolkit.org/7.15.html
	lookup	false	true	http://id.loc.gov/ontologies/bibframe/colorContent	Color Content	http://access.rdatoolkit.org/7.17.html
	resource	false	true	http://id.loc.gov/ontologies/bibframe/supplementaryContent	Supplementary Content	http://access.rdatoolkit.org/7.16.html
	resource	false	true	http://id.loc.gov/ontologies/bflc/relationship	Related Works	http://access.rdatoolkit.org/rdachp25_rda25-65.html
	resource	false	true	http://id.loc.gov/ontologies/bibframe/hasExpression	Related Expressions	http://access.rdatoolkit.org/rdachp26_rda26-25.html

Fig. 2. UAL Monographs profile extract. (Image by Bigelow and Sparling 2020)

Given the importance of RDA for PCC, past work was leveraged for the creation of UAL Continuing Resource and Monographs application profiles. In particular, the mappings from CSR (Balster, Rendall, and Shrader 2018) and BSR (BIBCO Mapping BSR to BIBFRAME 2.0 Group 2017) to BIBFRAME provided a quick reference to ensure that Sinopia application profiles captured key elements of description. This initial launch point was then informed by iterative phases of development and feedback with cataloguers at UAL and collaboration with others in LD4P2. The results are still a work in progress, but we now have functional application profiles that demonstrate an implementation scenario for RDA in linked data with BIBFRAME.

JLIS.it

The creation of a functioning linked data editor through LD4P2 was very impactful, so again it is important to ask what the problems are in terms of bibliographic control. Overall the challenges here are tied to the successes. As we have referred to beautiful chaos, necessary innovation to support linked data implementation, almost by definition must go beyond current infrastructure for standards development. With multiple concurrent projects and implementations and no single standards body guiding shared practice, slightly different approaches have emerged. On the other hand, theories and practices have been confirmed where multiple communities have come to the same conclusion based on independent analysis, as with the emergence of the svde:Opus and bflc:Hub in close comparison with the LRM Work.

Convergence: The Opus

One key difference between RDA and BIBFRAME that surfaces in much of the literature is the differentiation between core classes (RDA: Work/Expression/Manifestation/Item; BIBFRAME: Work/Instance/Item). In BIBFRAME the use of bf:hasExpression and bf:expressionOf helps solve this, but ultimately this ends up as a Work-Work relationship and the impact of which has been a matter of considerable discussion (Heuvelmann 2018). Happily, work in the Share-VDE community and at LC has attempted to address this discrepancy with BIBFRAME extensions.

In 2018 the Share-VDE Work ID Working Group (now called the Sapientia Entity Identification Working Group) was formed with the initial charge to review the creation of works and work identifiers for BIBFRAME data converted from MARC by Share-VDE. This in itself was a key project to support interchange by developing universal identifiers for works, but through the analysis of data sets from participating libraries the Working Group identified two key finding:

1. While Work → Expression relationships can currently be expressed in BIBFRAME, these are ultimately Work-Work relationships, and determining the initial or primary work, or hierarchical relationships between works may prove difficult with this structure.
2. Through conversion from MARC to BIBFRAME, or automatic work ID generation based on BIBFRAME elements, unless we can define a difference (a fingerprint for each cluster or constellation) between Work and SuperWork [renamed as Opus] elements then these relationships (work-expression) cannot be captured through conversion or automated processing. With the scale of data conversion underway, not doing this would seem like a missed opportunity. Once a separate fingerprint is defined for this primary work, it needs a name, thus the creation of SuperWork [Opus]. (Bigelow 2019)

Following these initial findings in 2018, the svde:Opus was developed in relation to the svde:Work based on iterative analysis of library collections converted from MARC to BIBFRAME and utilizing LRM and RDA elements as a guide. The model that surfaced (see Figure 3) with the svde:Opus as a type of bf:Work performs something of an ontological magic trick, preserving core elements and definitions for BIBFRAME for those that choose not to use the extension, but allowing for the benefits of the Opus and use with RDA.
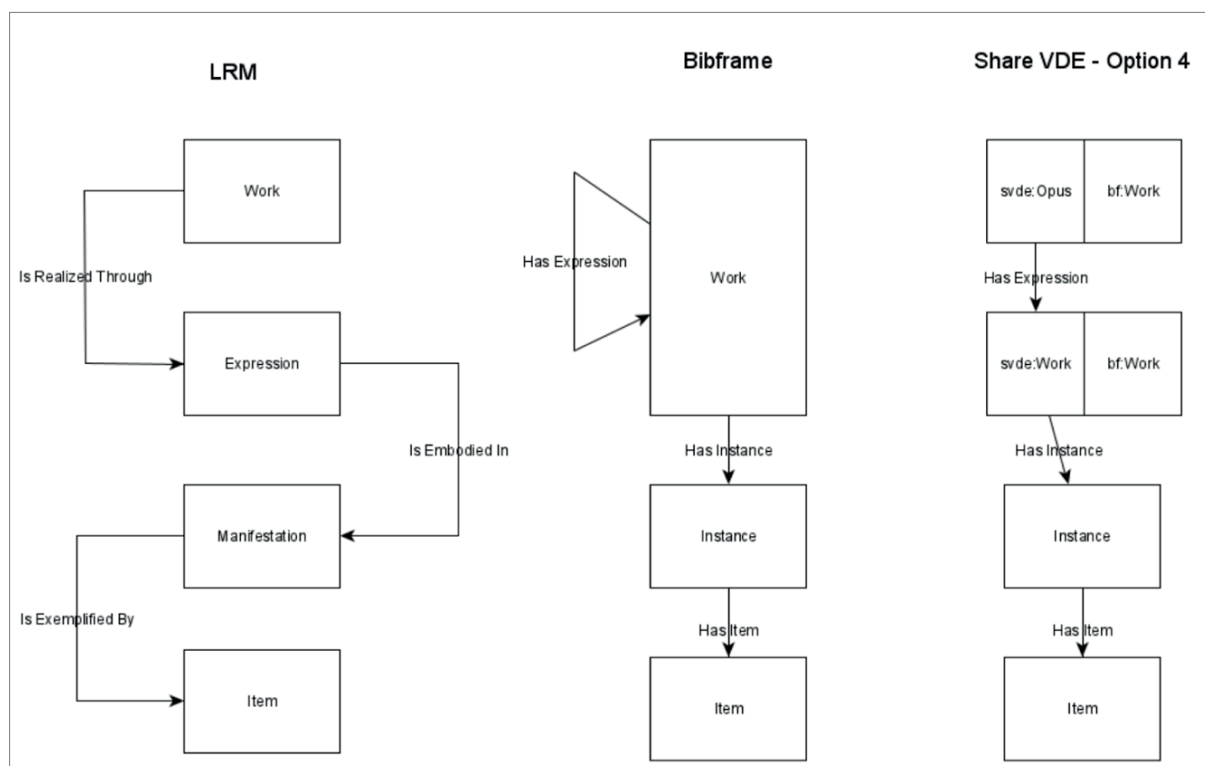
JLIS.it

Fig. 3. LRM, BIBFRAME and Share-VDE model comparison. (Ford, Kevin 2020b. [Share-VDE - Option 4]. Created for the Share-VDE Sapientia Entity Identification Working Group)

It is worth emphasizing that the svde:Opus emerged as a result of large scale testing of real world data. This is a beautiful example of theory being proven by practice, while at the same time high-lighting the nature of the collaborative work on the application of RDA in BIBFRAME.

In parallel with the svde:Opus, LC developed the bflc:Hub. In this the Hub was "Pursued be-cause [they] realized [they] were trying to do too much with bf:Work" (Ford 2020a). In many ways LC's use case was very similar to the need for the Opus, further validating a general need for this level of description and work aggregation. At the same time though, the Hub was defined slightly differently, conceptualized to be "Intentionally brief. Intentionally abstracted. Designed to ensure they are lightweight and maximally (re)usable" (Ibid.). While the Opus and Hub are both excit-ing developments, how these extensions inform the development of BIBFRAME as a standard remains to be determined.

As Share-VDE data is available for reuse in Sinopia, UAL has incorporated the Opus into our application profiles for resource description, allowing this structure to be tested and immediately put into use by our cataloguers when adding new Instance or Work descriptions for an existing Share-VDE Opus. Further refinements to how the Opus is incorporated in our application profiles will likely be needed, but being able to work with it in a cataloguing editor has made this much more real and hopefully will inform development of more standard best practice as PCC data has been converted from MARC to BIBFRAME and is now hosted by Share-VDE (Picknally and Bigelow 2020).

JLIS.it

Conformance and questions

As captured by the Task Group on PCC Sinopia Application Profiles "It is well known that there is no official mapping between BIBFRAME and RDA. The closest we have are the LC profiles and the BSR – and CSR – to BIBFRAME spreadsheets from some years ago, but none of these is "official" (PCC Task Group on Sinopia Application Profiles 2020, 9). The creation of "official" mappings should be a high priority for the RDA Steering Committee (RSC) to support RDA implementation scenarios in BIBFRAME, yet for the time being their absence does not mean that the data does not work.

An important piece of this discussion about what "official" RDA is stems from differing opinions on what RDA needs to be for particular communities of practice. The PCC Position Statement on RDA in August 2019 indicated that

> It is important to remember that RDA and RDA/RDF are two different things. RDA instructions will always be more applicable to traditional library resources than to newly emerging material types. We might also consider that given one of our goals for linked data is to communicate and consume data from beyond libraries, our RDF serialization might need to be more approachable than the complexity of RDA/RDF. As such and because we will probably be in a long-term transition away from MARC, PCC will continue to treat RDA as a loose content standard and participate in RDA/RDF and BIBFRAME discussions to assess our ideal linked data output. (Program for Cooperative Cataloging Policy Committee 2019, 3)

This distinction is tied to further developments of RDA 3R where increasingly efforts appear to focus on shifting RDA from a content to an encoding standard with RDA/RDF. Keeping in mind the PCC community context for Sinopia development it should not be surprising that UAL application profiles approached RDA implementation with a focus on using it as a content standard. This does not preclude the use of RDA/RDF in UAL profiles, but instead means that it can be applied along with BIBFRAME properties as needed.

Further stressing the difference in definition, in May 2020 the RSC released a discussion paper on RDA Conformance, indicating the required use of RDA/RDF and RDA constrained elements. The paper outlined that "A metadata statement is either conformant with RDA or it is not; there is no utility in the concept of partial conformance of a statement" (Dunshire 2020, 3). This statement suggests a shift in approach for RDA away from being an encoding scheme agnostic content standard. Given that PCC is not using RDA/RDF in this way, it indicates that PCC data (in MARC or BIB-FRAME) cannot be considered RDA conformant and thus not an implementation scenario.

Similarly, despite early concerns about the use of RDA constrained elements in a linked data environment, the 2020 discussion paper highlighted that "The unconstrained element set is not an integral part of RDA, and its use in metadata statements is not conformant with RDA" (Dunshire 2020, 2). In 2013 Alan Danskin captured the issue here well:

> An aspect of the linked data vision is that metadata can break down barriers, including those silos erected within the cultural heritage sector to meet the specific needs of museums, archives and libraries. Placing constraints on linked metadata elements is a barrier to reuse. For example, RDA Publisher's Name is an RDF property with domain manifestation. This is consistent with the FRBR model but it

makes the element unattractive to users or communities who do not perceive a need to distinguish between Work, Expression Manifestation and Item. It has taken some time for JSC [Joint Steering Committee] to understand these perspectives and from JSC's perspective an element set without FRBR cannot be RDA. (4)

It appears that since 2013 JSC has only become more firm in this siloed worldview. This is an unfortunate policy approach and strongly points to the need for further community collaboration on standards development. Nevertheless, as mappings are established between RDA constrained and unconstrained elements, ultimately what is important is semantic interoperability. If in order to implement RDA in BIBFRAME PCC or other communities of practice need to cease being conformant with RDA, so long as the resulting BIBFRAME data works for interchange the focus should be on further collaborative effort towards that end.

RDA/RDF or BIBFRAME

Reflecting back on Figure 1, the distinction between use of BIBFRAME versus RDA/RDF for encoding is an important one. If we end up with a large number of libraries using both then we will want to ensure interoperability and reuse of data between them. While RDA is certainly comparable to BIBFRAME, there are notable differences, for example with some elements having one to many or many to one relationship (McCallum and Williamschen 2019). Nevertheless, as demonstrated by work on Sinopia application profiles, core element sets can clearly be mapped and utilized from one to the other, and this should also support mappings for interchange, or indeed the use of both in a shared data set. Similarly, a Sinopia BIBFRAME application profile can readily incorporate both mappings to RDA instructions, and utilize RDA/RDF lookups when needed to utilize RDA vocabularies, just as Share-VDE has shown that RDA/RDF can be used to enrich BIBFRAME data (Hahn, Bigelow, and Possemato 2021).

The complexities of historical MARC data through conversion

While determining interactions between emergent library linked data standards are important for moving forward, we must also consider that as libraries move to BIBFRAME the majority of BIBFRAME descriptions will have started as MARC records. As such, some consistency is needed for the choices we make on how to convert MARC descriptions to BIBFRAME. Here we must consider where our data reflects both changes in practice as cataloguing standards have evolved, and where communities of practice have taken different approaches to resource description in MARC. As a result conversion processes from MARC to BIBFRAME face the challenge of accounting for myriad variations. Whether looking at the needs of an individual library, consortia, or library system, the changes in standards and local practices over time need to be addressed when converting to BIBFRAME. The work done by Share-VDE on MARC to BIBFRAME conversion is a prime example of this. Given membership from national and research libraries across North America and Europe, multiple languages and variations resulting from unique communities of practice need to be analysed and accounted for through conversion.

JLIS.it

One initial approach to work through this was to analyse the results of the conversion process by looking at converted records from 1985 and 2015 separately. Along with a more comprehensive analysis by Share-VDE members and Transformation Council, this assessment informed adjustments to the Share-VDE MARC to BIBFRAME conversion processes (Share-VDE Advisory Council 2018). It is important to note that handling some of these differences requires decisions, specific solutions, and sometimes compromises. An example of changing standards over time is the need to account for records with and without 33X fields (using GMD). Similarly, there have been different approaches across institutions and time for handling 7XX fields for related Opus, Work and Instance.
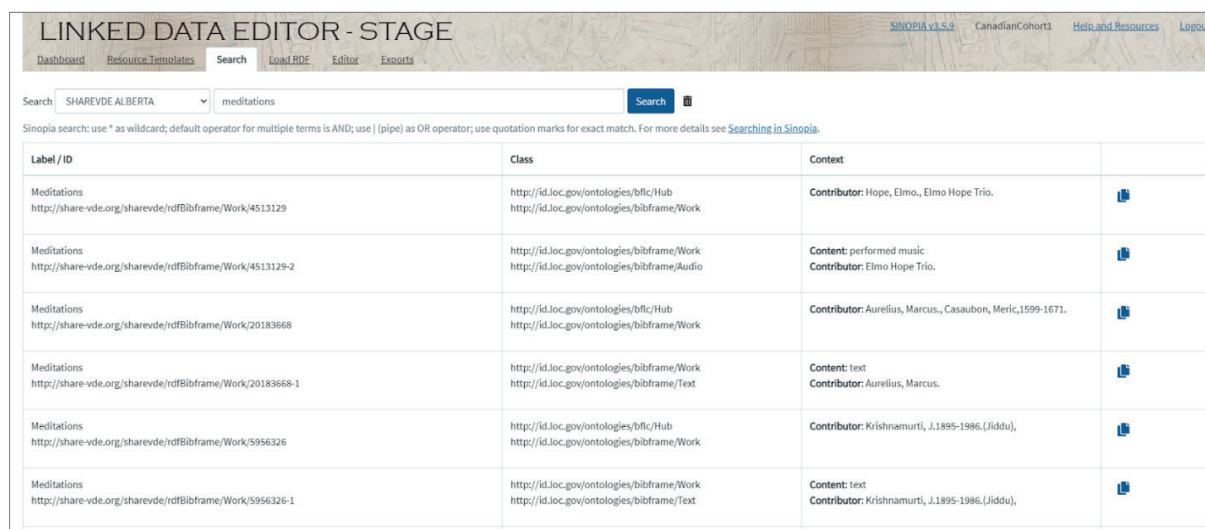
That many such variances need to be considered and decisions made for conversion, matching, and clustering again points to the desperate need for standardization, at least for core BIBFRAME elements. If these decisions are made independently for a given library or community for elements that are not solely local, then we are setting ourselves up for trouble as we begin sharing data (Park and Kipp 2019). Further, this speaks to the importance of transition planning. While MARC will need to be supported for some time to come, updates to it should be made with an awareness of the impact on multiple conversion processes.

Defining standard BIBFRAME data and infrastructure

Related to the issue noted above about decisions made for conversion from MARC to BIBFRAME, we also need to consider what the desired shape of BIBFRAME should be. It has been argued that "different interpretations derived from BIBFRAME's definitions, aiming to provide flexibility, may result in different implementations, hindering interoperability not just in mappings, but also between BIBFRAME implementations" (Zapounidou, Sfakakis, and Papatheodorou 2019, 301). To date we have encountered multiple examples of how different approaches to BIBFRAME modeling negatively impact data reuse. In order to support the transition from MARC to BIBFRAME and ensure data interoperability we require:

1. The data output of each MARC to BIBFRAME conversion process to be interoperable with the BIBFRAME created natively in RDF.
2. The ability to reuse BIBFRAME created in one community in other BIBFRAME stores.
3. BIBFRAME in various flavours to be converted to MARC with similar consistency.
4. New tools and processes to support various serializations of BIBFRAME (RDF XML, n-triples, n-quads, turtle, JSON-LD), or for the community to decide on which to use for development.

An example highlighting the need for points 1. and 2. is demonstrated through Sinopia copy cataloguing workflows. The Sinopia search feature allows users to search other sources for data reuse (currently BIBFRAME data created in Sinopia by other institutions and BIBFRAME data from the Share-VDE database). Figure 4 shows the results of a search for the UAL Share-VDE Work description of *Meditations*.

JLIS.it

Fig. 4. Screenshot of a search for a UAL Share-VDE Work description in the Sinopia editor

Reuse of BIBFRAME data in this way is a critical requirement for implementation, yet, because of the different choices made through the development path of Sinopia application profiles for original cataloguing in BIBFRAME and Share-VDE (where thus far BIBFRAME has been solely created through the process of conversion from MARC) challenges arose when attempting to import Share-VDE descriptions into Sinopia application profiles. Figure 5 illustrates how a number of triples from the Share-VDE description were unable to be brought into the PCC monographic work application profile.

Fig. 5. Screenshot of unused triples following the import of a Share-VDE Work description into the Sinopia PCC Monographic Work application profile in the Sinopia editor

In this case work is underway to resolve inconsistencies through collaborative effort with LD4P3, Share-VDE and PCC, but as more implementation cases emerge for BIBFRAME it makes sense to save work down the line by ensuring standardization to enable this kind of data reuse. An interesting note here is the continued lack of clarity on LC BIBFRAME data reuse outside of LC. LC is a member of PCC, and though one of the goals of LD4P3 is the creation of a shared PCC BIBFRAME datapool, there is little to indicate how LC will be contributing native BIBFRAME

descriptions. While standardized conversion from MARC does offer a pathway to consistent, reusable BIBFRAME data, the inability of native LC BIBFRAME to coincide with Share-VDE and Sinopia flavours of BIBFRAME supports the case for a swift standardization of a core BIBFRAME shape that works broadly for all libraries.

Addressing points 3. and 4. the case of conversion from BIBFRAME to MARC can be examined. In May 2020 LC released the XSLT for converting BIBFRAME to MARC along with associated conversion specifications (Library of Congress, n.d.b). Significant effort went into the MARC output, with LC knowing that MARC needed to be supported for many institutions for some time. As encouraging as this development is, in discussion on bibliographic control there are two challenges. The first issue is that BIBFRAME to MARC conversion output is dependent on the modeling choices and the resulting shape of the BIBFRAME that you start with. For example, you cannot successfully convert Sinopia BIBFRAME to MARC with the LC converter. This is a direct result of the differences in the Sinopia and LC application profiles which create different shapes of BIBFRAME. Similar inconsistencies in the shape of BIBFRAME and the impact on data interoperability are highlighted in the recently published *Final Report* of the PCC Task Group on Sinopia Application Profiles (2020). The second issue is that the LC converter only works with RDF/XML, while Sinopia uses JSON-LD and Share-VDE uses n-quads. These modelling differences and the need to utilize various serializations of RDF have the potential to encourage the development of new independent conversion processes which would add additional complexity when the goal is to standardize these processes.

Harnessing innovation and maintaining control

Throughout the course of BIBFRAME development and work across various communities on library linked data models, there have been calls for increased community engagement and the need for library linked data to be interoperable with data outside the library domain (Folsom 2020). As evidenced above though, it is equally pressing for real world library linked data to support interchange and interoperability between the institutions and projects creating, converting and publishing it. To do this there must be consensus on what constitutes standardized, core BIBFRAME data. To date, BIBFRAME development has been iterative, built initially by LC, but subsequently shaped by implementers through feedback provided to LC. Since the early days, LC has acknowledged that the BIBFRAME model,

> like MARC, must be able to accommodate any number of content models and specific implementations, but still enable data exchange between libraries. It needs to support new metadata rules and content standards that emerge, including the newest library content standard - RDA (Resource Description & Access). The BIBFRAME model must therefore both broaden and narrow the format universe for exchange of bibliographic data. (Miller et al. 2012, 5)

Community efforts and experimentation utilizing BIBFRAME have demonstrated its ability to broaden our universe. Experimentation has led to the creation of unique community extensions, format specific application profiles, and mappings between other emergent and project-specific library linked data models. It has also allowed us to work together as a library community, sup-

JLIS.it

ported by project partners, to begin building the systems and infrastructure we need to start converting, creating, editing, and making BIBFRAME data discoverable to our users. However, to support a working BIBFRAME data ecosystem, we now need to narrow our focus and define our core standards to support BIBFRAME interchange and conversion to maintain control across implementations. Moreover, the process of BIBFRAME implementation without exception requires a period where hybrid systems are in place (utilizing both BIBFRAME and MARC). This complex ecosystem requires standard practice more than we have ever needed it.

Experimentation and iterative development is a common characteristic of ontology building in LAM domains (Desmeules, Turp, and Senior 2020) and BIBFRAME is no exception. In fact, as noted, the BIBFRAME model's flexibility in implementation (Zapounidou, Sfakakis, and Papatheodorou 2019), while allowing for exploration and extensions across multiple communities, has led us to an impasse if we want to move ahead with wide implementation. With this knowledge, how do we move forward and define standards for BIBFRAME that support creation, reuse and conversion workflows? To do so we argue the following conditions need to be met:

1. Define core BIBFRAME elements necessary for resource description
    Defining core BIBFRAME elements is needed to facilitate the creation, reuse and conversion of BIBFRAME data between libraries. It is noteworthy, then, that PCC specific application profiles developed by the Task Group on PCC Sinopia Application Profiles were released alongside their final report in November 2020. The report outlines that

    > The intention of these templates is to provide a structured core of resource templates that allow catalogers to create PCC-level descriptions with uniform modeling and a basic set of vocabularies. It is hoped that they serve as the basis for a formal PCC standard (as an extension to the current BSR and CSR) at some point, and that in feeding the PCC data pool, serve as a pool of well-structured data to share, and provide vendors and developers data with which to experiment. (PCC Task Group on Sinopia Application Profiles 2020, 3)

    This is an excellent start towards standardization for the PCC community and hopefully it will extend to other communities and institutions. These application profiles support the identification of core BIBFRAME elements with attention to RDA implementation within them. They will also provide a template through which to test the resulting data. They will not, however, resolve the inconsistencies between the shape of BIBFRAME data being created and shared by other sources, such as Share-VDE and LC.

2. Define a standard BIBFRAME model and "shape" to support conversion and data reuse
    Data modelling assessment within the LAM domain has been shown to be an often ambiguous task (Desmeules, Turp, and Senior 2020). In particular we know that challenges often arise around implementing the data model and sample data in a technical production environment in order to assess it's success (Ibid.). To date, the complexity of building systems to support the use and analysis of BIBFRAME data has been a barrier to effective evaluation of the ideal "shape" of BIBFRAME to support LRM user tasks. However, with the data stores and discovery systems being developed by Share-VDE and LD4P, we are

JLIS.it

now in a position to use BIBFRAME's flexibility to our advantage to iterate and test standard BIBFRAME core application profiles to verify their utility for cataloguers and users alike. Once a BIBFRAME core and model are defined and tested, cataloguers and tool developers can create with confidence knowing their work will have wide application.

3. Define MARC use cases in a BIBFRAME environment
An interesting nuance of the discussion around BIBFRAME standardization is the need to determine use cases and standards to cover what we expect from MARC that has been converted from BIBFRAME. One approach (as represented by the LC converter) is to continue supporting MARC interchange for use in discovery. Another alternative approach could be to utilize BIBFRAME descriptions for discovery purposes, but utilize a much simpler, slim MARC output for inventory control in existing MARC systems. The later approach could simplify conversion processes for libraries moving to BIBFRAME, but would have obvious implications for metadata reuse. Further investigation into these points is timely as LD4P3 is currently developing separate BIBFRAME to MARC processes to support the conversion of native Sinopia BIBFRAME data.

4. Define implementation scenarios for the use of RDA 3R in BIBFRAME
Along with defining BIBFRAME standards, there is also the need to determine how the larger cataloguing community will be implementing RDA 3R in BIBFRAME to insure data interoperability and reuse. Similarly, where RDA/RDF is utilized independent from BIBFRAME clear mappings should be a priority to ensure interoperability and support use cases for data reuse.

5. Develop and coordinate implementation timelines for both RDA and BIBFRAME
Implementation timelines are necessary to make clear when both standards will be supported for application and exchange. When timelines are in place, libraries will be able to make more informed decisions about local practice and investments in transition.

Finally, wider community initiatives, best practices, and feedback loops need to continue to develop in order to successfully begin BIBFRAME implementations with a focus on bibliographic control. We have seen the start of a library community of practice around linked data with the establishment of the LD4 Community. The recent recommendations from the PCC Task Group on Sinopia Application Profiles (2020) that the PCC establish workflows for metadata reuse and investigate interoperability with the Share-VDE data model are also promising steps forward for bibliographic control within BIBFRAME. While welcome developments, it is also necessary to create open feedback loops between LC, other large scale projects and BIBFRAME implementers, and to establish relationships with the wider linked data community (Folsom 2020) to develop a BIBFRAME model and supporting systems that will enable bibliographic control. Here prioritizing transparency around ongoing and future developments to the BIBFRAME ontology and technical infrastructure (along with supporting analyses and user testing data) will be necessary to ensure BIBFRAME implementers can move forward on a shared path.

JLIS.it

All of these steps to maintaining bibliographic control in a BIBFRAME environment point to the need for community wide planning, standardization, and transparent communication. As always, innovation will still be necessary to ensure projects move forward in a way that serves libraries and library users, while leveraging the new systems and discovery potential linked data affords. Supporting the basic needs of interoperability through the refinement of a standardized BIBFRAME core will provide the library community with a solid foundation on which to build and facilitate the process of harnessing innovation for wider application.

References

Axelsson, Peter. 2018. "KB Becomes the First National Library to Fully Transition to Linked Data." National Library of Sweden: My News Desk, last modified July 5, accessed April 12, 2021, https://www.mynewsdesk.com/se/kungliga_biblioteket/pressreleases/kb-becomes-the-first-national-library-to-fully-transition-to-linked-data-2573975.pdf.

Baker, Thomas, Karen Coyle, and Sean Petiya. 2014. "Multi-Entity Models of Resource Description in the Semantic Web: A Comparison of FRBR, RDA and BIBFRAME." *Library Hi Tech* 32 (4): 562-582. doi:10.1108/LHT-08-2014-0081.

Balster, Kevin, Robert Rendall, and Tina Shrader. 2018. "Linked Serial Data: Mapping the CONSER Standard Record to BIBFRAME." *Cataloging and Classification Quarterly* 56 (2-3): 251-261. doi:10.1080/01639374.2017.1364316.

BIBCO Mapping BSR to BIBFRAME 2.0 Group. 2017. *Final Report to the PCC Oversight Group.* https://www.loc.gov/aba/pcc/bibframe/TaskGroups/BSR-PDF/FinalReportBIBCO-BIB-FRAME-TG.pdf.

Bigelow, Ian. September 2019. "Opus Ex Machina: Modelling SuperWork and Work Entities in BIBFRAME." Presentation at the 3rd Annual BIBFRAME Workshop in Europe. Stockholm, Sweden. https://www.kb.se/download/18.d0e4d5b16cd18f600eacb/1569309579935/Opus%20Ex%20Machina%20-%20Present.pdf

Bigelow, Ian, and Abigail Sparling. "Control or Chaos: Embracing Change and Harnessing Innovation in an Ecosystem of Shared Bibliographic Data". Firenze, Italy, 2021. https://www.youtube.com/embed/ybUDrILt0kI?start=7714&end=12441.

Bigelow, Ian, Danoosh Davoodi, Sharon Farnel and Abigail Sparling. August 2018. "Who Will be Our bf: Comparing techniques for conversion from MARC to BIBFRAME". Paper presented at IFLA WLIC, Kuala Lumpur, Malaysia. http://library.ifla.org/2194/.

Cronin, Christopher. January 2011. "Will RDA Mean the Death Of MARC?: The Need For Transformational Change to Our Metadata Infrastructures." Presentation at the American Library Association Midwinter Meeting, San Diego, CA. https://www.academia.edu/1679819/Will_RDA_Mean_the_Death_of_MARC_The_Need_for_Transformational_Change_to_our_Metadata_Infrastructures

JLIS.it

Danskin, Alan. 2013. "Linked and Open Data: RDA and Bibliographic Control." *JLIS.It* 4 (1): 147-160. doi:10.4403/jlis.it-5463.

Desmeules, Robin Elizabeth, Clara Turp, and Andrew Senior. 2020. "Exploring Methods for Linked Data Model Evaluation in Practice." Journal of Library Metadata 20 (1):65-89. doi:10.1080/19386389.2020.1742434.

Dunshire, Gordon. 2020. "RDA Conformance: Discussion paper for RSC." RDA Steering Committee. http://www.rda-rsc.org/sites/all/files/RDA%20conformance%20proposal.pdf.

Farnel, Sharon, Ian Bigelow, Denise Koufogiannakis, Leah Vanderjagt, Geoff Harder, Peter Binkley, and Sandra Shores. "Moving Forward with Linked Data at the University of Alberta Libraries". (Edmonton: University of Alberta Library, 2018) https://docs.google.com/document/d/1t-5Kh-n3Ctbl9HcXHr4_cDneNzukaaAnm-AXh-ixUY0E/edit.

Ford, Kevin. 2020a. "On Bibframe Hubs." Presentation at the American Library Association Midwinter Meeting, Philadelphia, PA. https://docs.google.com/presentation/d/1oPEMnnpMnCIXo-IMWztThm_XGc4lEzaoP/edit#slide=id.p1.

———. 2020b. "[Share-VDE - Option 4]". Share-VDE Sapientia Entity Identification Working Group.

Folsom, Steven M. 2020. "Using the Program for Cooperative Cataloging's Past and Present to Project a Linked Data Future." *Cataloging & Classification Quarterly* 58 (3/4): 464–71. doi:10.1080/01639374.2019.1706680.

Guerrini, Mauro and Tiziana Possemato. 2016. "From Record Management to Data Management: RDA and New Application Models BIBFRAME, RIMMF, and OliSuite/WeCat." *Cataloging and Classification Quarterly* 54 (3): 179-199. doi:10.1080/01639374.2016.1144667.

Hahn, Jim, Ian Bigelow, and Tiziana Possemato. March 2021. "Share-VDE Model Overview: SEI-WG Update for the Share-VDE Virtual Workshop." Presentation at Share-VDE Virtual Workshop. https://docs.google.com/presentation/d/116PwHecnqooEjW3c4Eks77Ah6RilpiOpAfCe61K4UoI/edit#slide=id.gbfe20d43fa_0_0

Heuvelmann, Reinhold. September 2018. "RDA / MARC / BIBFRAME: some observations" Presentation at the European BIBFRAME Workshop, Fiesole, Italy. https://www.casalini.it/EBW2018/web_content/2018/presentations/Heuvelmann.pdf.

IFLA Study Group on the Functional Requirements for Bibliographic Records, and Standing Committee of the IFLA Section on Cataloguing. 1998. *Functional Requirements for Bibliographic Records Final Report*. Berlin: De Gruyter. doi:10.1515/9783110962451.

International Federation of Library Associations and Institutions. 2017. "Best Practice for National Bibliographic Agencies in a Digital Age: Bibliographic control." https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8911.

Lendvay, Miklós. September 2020. "Hungarian National Library Platform Implementation." Presentation at BIBFRAME Workshop in Europe Virtual Event. https://www.casalini.it/bfwe2020/web_content/2020/presentations/lendvay.pdf.

JLIS.it

Library of Congress. n.d.a. "BIBFRAME 2.0 Implementation Register." Library of Congress., last modified n.d., accessed April 13, 2021, https://www.loc.gov/bibframe/implementation/register.html.

———. n.d.b. "New BIBFRAME-to-MARC Conversion Tools." Library of Congress., last modified n.d., accessed January 20, 2021, https://www.loc.gov/bibframe/news/bibframe-to-marc-conversion.html.

Lionette, Anna. 2021. "Share-VDE institutions." Share-VDE Wiki, https://wiki.share-vde.org/wiki/ShareVDE:Main_Page/SVDE_institutions.

Magda El-Sherbini. 2018. "RDA Implementation and the Emergence of BIBFRAME." *JLIS.It* 9 (1):66-82. doi:10.4403/jlis.it-12443.

McCallum, Sally and Jodi Williamschen. September 2019. "RDA and BIBFRAME at the Library of Congress." Presentation at the European BIBFRAME Workshop, Stockholm, Sweden. https://www.kb.se/download/18.d0e4d5b16cd18f600eac6/1569246418227/LC_EUBF2019-RDA+BF.pdf.

McGrath, Kelley. January 2011. "Will RDA Kill MARC?" Presentation at the American Library Association Midwinter Meeting, San Diego, CA. https://pages.uoregon.edu/kelleym/publications/McGrath_Will_RDA_Kill_MARC.pdf

Miller, Eric, Uche Ogbuji, Victoria Mueller, and Kathy MacDougall. 2012. *Bibliographic Framework as a Web of data: Linked data model and supporting services*. Library of Congress. https://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.

Park, Hyoungjoo and Margaret Kipp. 2019. "Library Linked Data Models: Library Data in the Semantic Web." *Cataloging & Classification Quarterly* 57 (5) (July 4,): 261-277. doi:10.1080/01639374.2019.1641171.

PCC Task Group on Sinopia Application Profiles. 2020. *Final Report*. https://www.loc.gov/aba/pcc/taskgroup/Sinopia-Profiles-TG-Final-Report.pdf.

Picknally, Beth, and Ian Bigelow. August 2020. "PCC BIBFRAME data: PCC collaboration with Share-VDE." Presentation at the PCC At Large Virtual Meeting. https://www.loc.gov/aba/pcc/documents/Virtual-5-ShareVDE-PCC-2020.pdf

Program for Cooperative Cataloging Policy Committee. 2019. *PCC's Position Statement on RDA*. https://www.loc.gov/aba/pcc/rda/PCC%20RDA%20guidelines/PCC-Position-Statement-on-RDA.docx.

———. 2020. *Implementation of the New RDA Toolkit*. https://www.loc.gov/aba/pcc/documents/PoCo-2020/newRDA%20ImpementationPlanNov2.pdf.

RDA Steering Committee. 2020. "What You Should know about the December Switchover", RDA Toolkit, https://www.rdatoolkit.org/node/230.

Samples, Jacquie. January 2011. "Will RDA Mean the Death of MARC?" Presentation at the American Library Association Midwinter Meeting, San Diego, CA. https://connect.ala.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=0ca55eba-615b-4aa0-aa84-1fba223483d5

JLIS.it

Samples, Jacquie, and Ian Bigelow. 2020. "MARC to BIBFRAME: Converting the PCC to Linked Data." *Cataloging & Classification Quarterly* 58 (3/4): 403–17. doi:10.1080/01639374.2020.1751764.

Seikel, Michele and Thomas Steele. 2020. "Comparison of Key Entities within Bibliographic Conceptual Models and Implementations: Definitions, Evolution, and Relationships." *Library Resources & Technical Services* 64 (2): 62-71. doi:10.5860/lrts.64n2.62.

Share-VDE. 2019. "Share-VDE major achievements in 2019." https://drive.google.com/drive/search?q=share%20vde%20major%20achievements

Share-VDE Advisory Council. 2018. *SVDE recommendations to improve the MARC to BIBFRAME conversion.* https://drive.google.com/drive/folders/1PsVMvLEMXtzL8vKBmAbHeFt8FZcvOJs

Stanford University. 2018. "Stanford Libraries announces Linked Data for Production (LD4P) cohort members and subgrant recipients," Stanford Libraries. https://library.stanford.edu/node/155851.

Taniguchi, Shoichi. 2017. "Examining BIBFRAME 2.0 from the Viewpoint of RDA Metadata Schema." *Cataloging and Classification Quarterly* 55 (6): 387-412. doi:10.1080/01639374.2017.1322161.

Taniguchi, Shoichi. 2018. "Mapping and Merging of IFLA Library Reference Model and BIBFRAME 2.0." *Cataloging and Classification Quarterly* 56 (5-6): 427-454. doi:10.1080/01639374.2018.1501457.

Zapounidou, Sofia. 2020. "Study of Library Data Models in the Semantic Web Environment." Ionian University, Department of Archives, Library Science and Museology PhD Thesis. https://zenodo.org/record/4018523.

Zapounidou, Sofia, Michalis Sfakakis, and Christos Papatheodorou. 2019. "Mapping Derivative Relationships from RDA to BIBFRAME 2." *Cataloging & Classification Quarterly* 57 (5):278-308. doi:10.1080/01639374.2019.1650152.

JLIS.it

The multilingual challenge
in bibliographic description and access

Pat Riva[a]

a) Concordia University, http://orcid.org/0000-0001-6024-4320

Contact: Pat Riva, pat.riva@concordia.ca
Received: 15 April 2021; **Accepted:** 8 May 2021; **First Published:** 15 January 2022

ABSTRACT

Cataloguing has taken many steps towards greater internationalisation and inclusion, but one area remains stubbornly intractable: providing transparent access to users despite differences in language of descriptive cataloguing and language of subject access. As constructed according to present cataloguing practices, bibliographic records contain a number of language-dependent elements. This may be inevitable, but does not have to impede access to resources for a user searching in a language other than the language used for cataloguing. When catalogues are set up as multiple unilingual silos, the work of bridging the language barrier is pushed onto the user. Yet providing access through metadata is supposed to be the role of the catalogue. While a full theoretical approach to multilingual metadata is elusive, several pragmatic actions can be implemented to make language less of a barrier in searching and interpreting bibliographic data. Measures can be applied both in the creation of the metadata, and in adjusting the search. Authority control, linked authority files, and controlled vocabularies have an important part to play. Examples and approaches from the context of a newly established catalogue shared by a consortium of English language and French language university libraries in Québec, Canada.

KEYWORDS

Multilingual catalogues; bilingual cataloguing; bilingual publications; language of cataloguing; cross-linguistic subject searching.

JLIS.it

Universal Bibliographic Control

This international conference on *Bibliographic Control in the Digital Ecosystem* takes its context from the IFLA Professional Statement on Universal Bibliographic Control (UBC)[1] whose latest version was prepared by the IFLA Bibliography Section and endorsed in 2012.

In the original conception of UBC, first promoted in the 1970s (Anderson 1974), which was a very different technological context from today, the idea was for each national bibliographic agency (NBA) to create data for its own national publications once, while following standards to allow reuse of that data internationally. The idea was that by using the same form of access points, as established by the originating agency, it would be possible to exchange and integrate all the records into all the national catalogues. The focus was on efficiency and maximum sharing of effort.

However, global means multilingual. This concept of UBC did not take into account that users would have difficulty imagining the access points to use when these were devised in the language of cataloguing of the publishing country, not the user's preferred language, and that the number of different forms to search would increase depending on the origins of resources in the collection. As these access points can differ considerably, even without imagining the difficulties relating to different scripts, shifting this burden to the user is not compatible with our professional understanding of good service to the user. So in reality, NBAs could be informed by the work of their colleagues, but still needed to establish their own preferred forms and recatalogue resources to integrate them into their own catalogues. And this work falls less to NBAs than to their clients, libraries of all types around the world that collect materials published throughout the world.

And so the next conception of UBC, first proposed in the late 1990s, involved linking authority files contributed by different NBAs so that authority records describing the same entity but according to different choices of preferred language and script and different cataloguing conventions would be brought together via mapping (Tillett 2008). This is the thought that led to the Virtual International Authority File (VIAF) that we all know and use heavily (VIAF)[2]. And this is a powerful idea that translates nicely into the semantic web and linked open data (Willer and Dunsire 2013).

This still does not consider the display and retrieval of metadata, not only access points, from the user's point of view – a user who may be a multilingual.

User Need for Multilingual Access

All human beings unavoidably work in language, think in language. Language has a very deep effect on all we do. Arguably, we can do little with library resources without language to mediate our access. Even resources with primarily visual or auditory (non-linguistic) content are mediated via metadata that includes language, and writing systems.

As has been described (Riva 2020, 137-138), there are several layers of multilingualism. Many user communities are multilingual, library collections are multilingual, and individual users have a continuum of language ability in multiple languages, which is reflected in the resources they want

---

[1] https://repository.ifla.org/handle/123456789/448
[2] http://viaf.org

JLIS.it

to access. Multilingual is a perspective that can apply both to individual users and to the user community of a library as a whole.

Note that a person does not have to be perfect in a language to use library resources in that language. In many cases one can use a resource even without being able to read absolutely all of it. For example, the resource may itself be multilingual (consider facing page translations, or the proceedings of multilingual conferences), or the resource may have minimal text, such as some art catalogues, or maps or image collections.

Language of Cataloguing

A basic term in this discussion is *language of cataloguing.* It is a long-established term which seems to be considered obvious since it is never defined in the expected sources. It refers to the language used for all metadata, both descriptive and subject, that the cataloguer must provide in completing a resource description. This determines the linguistic suitability of the resulting record. A traditional assumption is that one catalogue will be built around one language of cataloguing.

RDA, Resource Description and Access[3], comes the closest to defining the concept in the definition of the principle of "Common usage or practice" found in the section on "Objectives and Principles governing RDA": "Data that are not transcribed from a manifestation that is being described should reflect common usage in the language and script chosen for recording the metadata."

RDA goes on to state: "An agent who creates the metadata may prefer one or more languages and scripts." RDA in its original formulation regularly, such as in instruction 0.11.2 *Language and Script*, used the carefully worded phrase "in *a* language and script preferred by the agency creating the data" [emphasis mine], not *the* preferred language of the agency, to explicitly allow for multilingual cataloguing agencies, but little is said about the practical consequences of having multiple preferred languages working together in a single catalogue. Common practices in this area have not yet emerged.

Catalogue Configurations

Despite considering the question for several years, the exact meaning of a multilingual or bilingual catalogue is still imprecise. The catalogue we want depends on what we think our users will need. Are we serving a population that only uses one language and minimally is interested in others? Then a traditional catalogue with a focus on a single language is best suited. All resources, regardless of the languages of their content (and to the extent that resources in these other languages are even collected), are described and accessed via one language.

Or is one library serving distinct sub-populations each with its own language and likely to be interested in only its own materials? Then a solution similar to the Library and Archives Canada *Bilingual Cataloguing Policy* (LAC 2003)[4], may suit. Under this policy, resources may be described once or twice, depending on the language of the content. Roughly speaking, French-language resources are described in French, English-language resources in English, and English-French bilin-

---

3 https://access.rdatoolkit.org/
4 https://www.bac-lac.gc.ca/eng/services/cataloguing-metadata/Pages/bilingual-cataloguing-policy.aspx

JLIS.it

gual resources in both languages, using two records. Although the available text of the policy is not yet updated since LAC's adoption of RDA, the determination for monographs of which treatment applies to a resource shows the details that must be considered in operationalizing this policy:

> Monographs
>
> 1. All French-language publications (including multilingual publications containing a substantial portion of text in French) will be catalogued in French, according to the Règles de catalogage anglo-américaines, deuxième édition, révision de 1998 and its updates. Subject headings will be assigned in both French and English.
>
> 2. All publications in other languages (i.e. those containing no substantial portion of text in French) will be catalogued in English, according to the Anglo-American Cataloguing Rules, second edition, 2002 revision and its updates. Subject headings will be assigned in both English and French.
>
> 3. All bilingual and multilingual publications containing substantial portions of text in both English and French will be catalogued twice, once in English and once in French. English subject headings will be assigned to the English record; French subject headings will be assigned to the French record.
>
> 4. Texts in Latin and instructional materials will be catalogued according to the language of the intended audience (i.e. those intended for a French-speaking audience will be catalogued in French; those intended for an English-speaking or other language audience will be catalogued in English). Subject headings will be assigned according to the policy outlined above at 1-3."

As a result, even if using the same catalogue, users interact primarily with metadata curated to be appropriate to the users' chosen language. However, it does create language silos, as users are guided to discovering only those resources in that language. This separation can be implemented using multiple catalogues, a solution that might make a lot of sense when the languages are in distinct writing systems. Scalability may become a concern under these approaches with the addition of more languages.

Grounding in Local Context

For another population, with individuals actively using multiple languages, the goal is to allow users to search once in the language of their choice and retrieve relevant material regardless of the materials' language. This is the use case of interest for the partnership of Quebec university libraries. Canada is bilingual federally, but the official language of Quebec is French. Of the 18 universities in Quebec, 15 use French as a language of instruction, and three teach in English. All the libraries catalogue in the language of instruction of the respective university, but collect in both English and French (and in many other languages depending on the programs of instruction that are offered). The partnership's combined user population includes a whole spectrum of English-French speakers, including scholars with reading knowledge of languages, and many international students, immigrants and first-generation Canadians. Thus the partnership catalogue must bridge this language gap for the user, at least for English-French bilingualism. Bilingual services were a major element taken into consideration in the design of the Sofia[5] catalogues that were launched in summer 2020 following two years of preparation.

---

[5] https://sofia-biblios-uni-qc.org/fr/

JLIS.it

User Display

Once the user has framed a search and retrieved records, the results need to be displayed to the user in a way consistent with the linguistic presentation of the interface. Is the content of the record adaptable to be appropriate to the user's language preference? One strategy to adapt the catalogue data is to store a single record and transform it so as to display according to the desired language. This seems like a natural extension of how the language of the user interface of a system, or of a website, can be switched between languages by a user. An easy part of the metadata to transform from one language to another is any value that is taken from a simple value vocabulary or controlled list. As long as display labels for those values exist in the user's desired language, the code can be displayed using its equivalent label in that language. Using codes is simple, cost-effective, and scalable (Aliverti 2019, 18). This is another point in favour of using controlled terms and established value vocabularies as much as possible, and it has the added benefit of being easier to adopt in a linked data context, using the mechanism of preferred labels with associated language attributes (Willer and Dunsire 2013, 182-192).

Preferred Forms of Names and Role the Authority File

In addition to showing appropriate display labels for controlled terms and coded values, the forms of access points displayed to the user should be language-appropriate. This is necessary because language affects the choice of form of name in some cases: "Choose a well established name in a preferred language" is the usual phrase. This affects the choice of name for classical authors, for example of Plato (Platone, Platon, etc.), and also personal names that include cataloguer-supplied elements, such as Popes, Saints, Sovereigns, etc.

With corporate bodies, the choice of language of name affects the preferred access points for international bodies (United Nations vs Nations Unies, etc.). It also affects government subdivisions, since the name of the country will have an established form in the preferred language of the cataloguing agency, but usually the sub-body will only have an established name in the language of the country of the body. A typical example of the resulting bilingual construction is the English language form for the Italian meteorological service, established in the PCC-NACO authority file as:

> 110   1_ |a Italy. |b Servizio meteorologico (n 2004021837)

An even more extreme example, for the office of the scientific attaché of the Italian Embassy in Belgium, in the form from the PCC-NACO authority file and suitable for an English-language agency, displays three languages. The name of the country, Italy, is in English, as is the qualifier for the country where the Embassy is found, Belgium. The term for an embassy is given in Italian, the language of the body, but the language of the specific office is in French, the language of the name of that body as used in Belgium.

> 110   1_ |a Italy. |b Ambasciata (Belgium). |b Bureau de l'attaché scientifique (n 2004120329)

JLIS.it

Displaying the language-appropriate forms of corporate bodies such as these examples, or other language-dependent names, requires maintaining equivalencies for each of the languages being supported. Creating a single authority file holding preferred forms in several languages within each record is the approach selected in several multilingual national libraries. Cohen describes the National Library of Israel's name authority file (Cohen 2020) which includes forms appropriate in English, Hebrew, Arabic and Russian, each in the relevant script. The Swiss Library (Lehtinen and Clavel-Merrin 1998) also describes an approach with multiple preferred forms stored in repeatable fields in a single authority record. As explained by Aliverti (Aliverti 2019, 22-24), a machine can only match a recorded name to a language if the language corresponding to the name is explicitly coded. In both these cases, the authority file is under the control of a single agency, and although multiple languages are used, there is a small, established list of the languages that are supported. Scaling these approaches to ever more languages would have significant costs.

By linking authority records contributed independently from different authority files with different languages of cataloguing, it should be possible for a system to look up an entity and retrieve an appropriate form in the desired language to display with the bibliographic metadata. Selecting linguistically appropriate display forms from sets of authority records for the same entity is the exact issue that VIAF was designed to solve. So far VIAF has remained a cataloguer's tool and is not yet implemented as widely as it could be in interfaces for end-users.

But there is more to the catalogue and its data than access points from the name authority file. This brings us to consider the languages used in the description.

Multiple Shared Records

The approach of taking multiple records and linking them, instead of transforming a single record for display, can be applied to bibliographic records as well as to authority records. Then instead of manipulating the data elements within a single record, the whole record that corresponds to the user's desired language is selected for display. A single cataloguing agency applying stable cataloguing practices in its own catalogue can control the linkage between different language of cataloguing records for the same resource, thus ensuring equivalent service to each language group. On the other hand, sharing the work among different agencies, as in the Sofia catalogue, means pulling together metadata contributed by different agencies, each working independently in its own language of cataloguing. Then the question shifts to one of recognition that the different records describe the same resource. This recognition depends on standards and their consistent application in a shared environment, something libraries have considerable experience with, but the community working together must broaden in size to cover multiple languages.

How can that link be made? There is not yet a MARC 21 field that can serve to hard-link two descriptions for the same resource that are parallel language descriptions. Standard identifiers for the resource can be a start. Recording the identifiers is objective and should not be dependent on any of the cataloguing agencies involved. Also external to the metadata is any transcribed data from the resource itself, if selected and recorded consistently. And so these manifestation statements serve as an identifying element for the manifestation.

JLIS.it

Language in Description

Much of the descriptive metadata depends on the language of the resource, or at least the language of the resource's identifying information. This data is a surrogate for the resource and not to be transformed for display. All transcribed data – the manifestation statements – depends on the language used in the resource: title proper, statements of responsibility, edition, series. As do notes quoted from the resource. Although, in some cases this data does not reflect the language of the content, usually it does.

In contrast, there are a number of places in the descriptive portion of a record which depend on the language of cataloguing. Present in almost all descriptions:

- Cataloguer-supplied notes: since the cataloguer must compose them, this needs to be done in a language the cataloguer is competent to write in.
- Qualifiers: such as for ISBNs, other standard numbers.
- Prescribed terms: such as in physical description, there are many such terms, all over the description.

More infrequent situations:

- Supplied title proper: when there is no title proper and the cataloguer must devise a title, this is generally in the language of the catalogue.
- Choice of title page for multilingual publications: in certain contexts, the language of cataloguing plays a determining role.

Examples of Standard Multilingual Publications

The choice of a source of information has considerable impact on the resulting description. Some multilingual publications also present parallel titles and other data in one or more sources.

A first case is illustrated by the Canadian Modern Languages Review = La revue canadienne des langues vivantes (figure 1). The source clearly presents two parallel titles. Bibliographic data is in both languages, but presented alternately on a single source. Following the normal left-to-right conventions, there is no doubt that the title to the left, the English title, should be transcribed first as the title proper. This decision is not dependent on the language of cataloguing. Since this is a journal, the content includes articles in one or the other of the languages, but only editorially supplied content is in both languages.

A slightly more complex case is presented by the proceedings of the IFLA International Meeting of Experts for an International Cataloguing Code 5 (figure 2). It has three parallel titles, in English, French and Portuguese, on the same source, which by convention the cataloguer will read from top to bottom, again resulting in the choice of the English parallel title as the title proper, regardless of language of cataloguing. Contributions are mainly translated into all three languages.
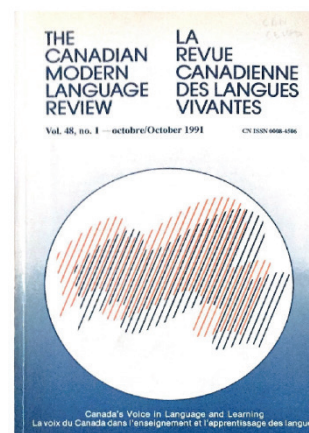

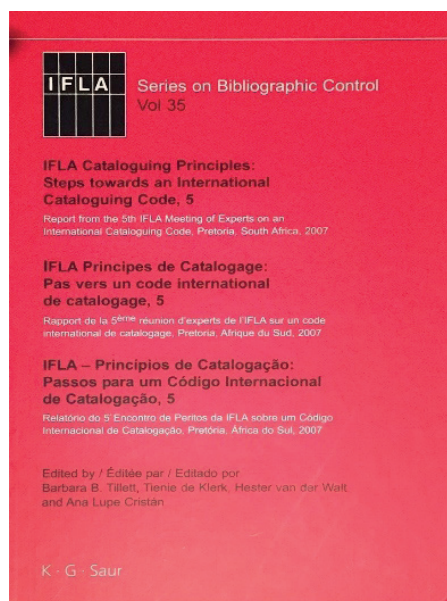Fig. 1. Canadian Modern Languages Review
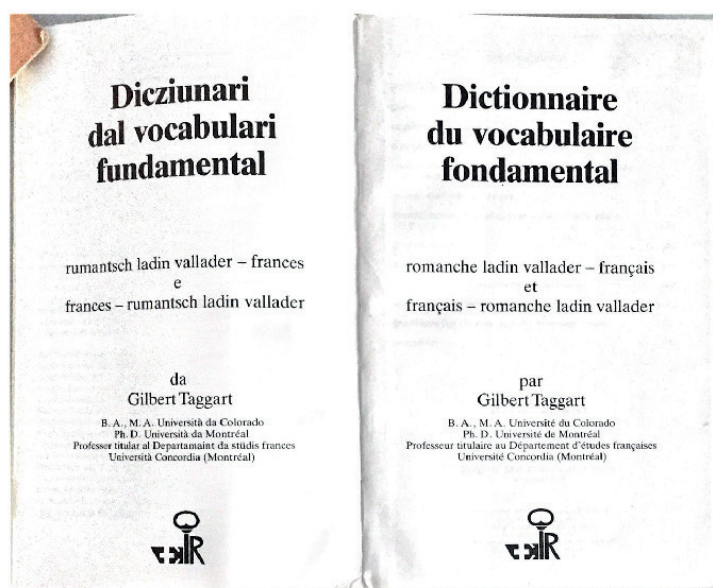
JLIS.it

Fig. 2. IFLA Cataloguing Principles

Fig. 3. Rumantsch-French bilingual dictionary

In this bilingual Rumantsch-French dictionary (figure 3) there are two full title pages. Applying the left-to-right convention again, this time to the choice of title page, the cataloguer is clearly directed to record the Rumantsch title first, as the title proper. The content of the dictionary alternates between the two languages.

In these three cases, since the same title proper will be chosen regardless of the language of cataloguing, the identification of the resource will be constant and there is a good chance that algorithms can match records catalogued in different languages as being for the same resource.

Paradox of Tête-Bêche Publications

For one type of publication, the normally evident decision about source of information is anything but. Consider the tête-bêche publication layout. This is variously described as head-to-toe or text on inverted pages. It is usually used for relatively short technical or government reports for bilingual corporate bodies or jurisdictions. It is a very specialized publication format limited by its physical characteristics to two languages of text.

An example is Excursion B-19. The construction is best seen when the booklet is opened flat so that both covers can be seen at once (figure 4). The two covers both look like front covers, but presented on inverted pages. Text runs from each cover to meet in the middle. Opening the booklet from the English cover reveals the English title page (figure 5), while turning the booklet to open it from the French cover reveals the French title page (figure 6). There are two front covers and two title pages that are of exactly equal prominence. There is no physical distinction, or right way up! Each language is treated exactly equally. Is there any objective way one of these title pages can be said to be first? No! The choice of title page is arbitrary. With no characteristic inherent in the

publication to guide the cataloguer's choice, the criterion that remains is the language of cataloguing. For these publications, cataloguing conventions direct the cataloguer to choose the title page matching the language of cataloguing. Yet the publication can still be described as a whole, much as any facing-page translation or the bilingual dictionary with two adjacent title pages, by giving the title from the title page not chosen as a title from added title page.
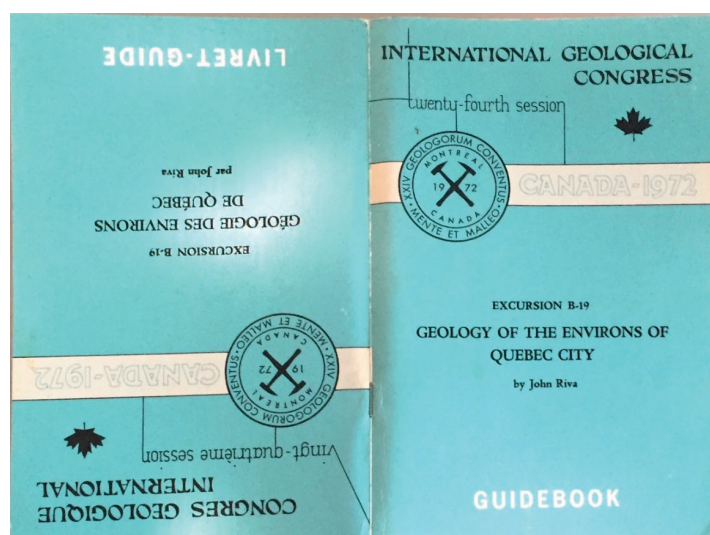

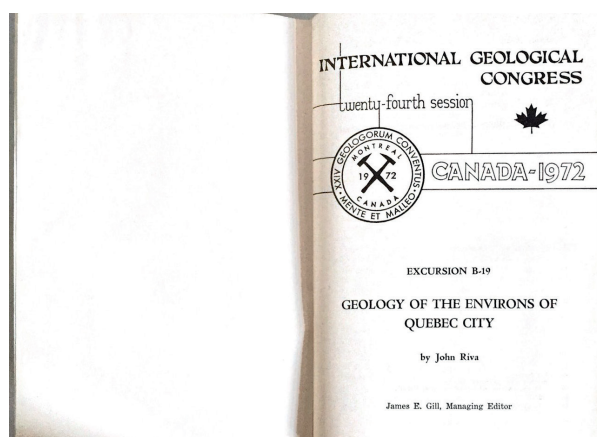Fig. 4. Tête-bêche publication open to show both covers

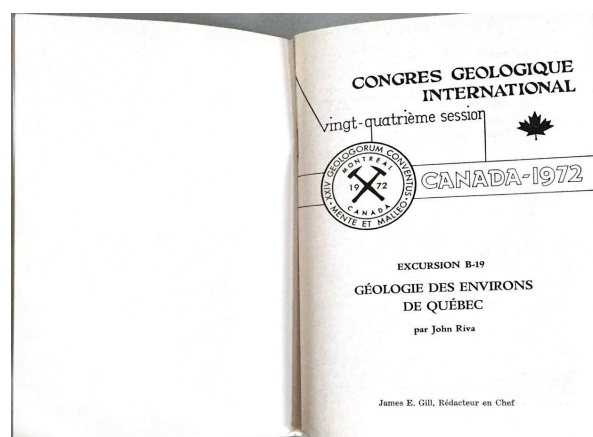Fig. 5. English title page of tête-bêche publication

Fig. 6. French title page of tête-bêche publication

This results in two records that differ in many ways based on the language of cataloguing. Although the two records present the resource fully appropriately according the language chosen for cataloguing, and serve users well, it seems unlikely that these record pairs can be machine-detected as being language of cataloguing variants for the same resource. The choice of title page has affected the choice of title proper, all other transcribed statements, pagination and probably many

other subtle details. Unless there is a standard number (note that Excursion B-19 does not have an ISBN), it would be difficult for an algorithm to match these parallel records, yet distinguish a tête-bêche publication from the entirely different case of two records in different languages of cataloguing that represent different language expressions that are not issued bound together. This is where a cataloguer-assigned link between records would be convenient, to allow overriding of the apparent differences.

Another particularity of the tête-bêche publication is what happens when it is digitized. The digitization has to start at one cover and linearly scan the document. Generally, the inversion is not preserved and the two expressions are scanned consecutively by returning to the other cover once the centre is reached. Because of the file layout, the choice of title page is forced according to whichever language is presented first in the file. In digital form, the choice of title proper is not dependent on language of cataloguing and the resource can be catalogued in the same manner as any bilingual publication presented sequentially. The cataloguing is much easier, but now a new difficulty arises. Matching the digital reproduction to the original, even when both records use the same language of cataloguing, needs to rely on a linking field.

Topical Subjects and Classification

Strategies to provide subject access cross-linguistically have seen a lot of attention (Park 2007) and my aim here is not to provide a comprehensive review of that literature. Classification is enticing as a language switching hub, because the classification notations may appear to be language-neutral, but there are cultural expectations built-in to the design of classification, as basic as what topics go together, and which do not. Despite all this, a common classification can still be useful in facilitating multilingual rendering of resource metadata, by linking the classification notation to captions in different languages for display, as is done in the *Swiss Book*, the national bibliography of Switzerland, which uses captions for its Dewey Decimal Classification (DDC) subject categories in German, French, Italian, Rumantsch, and English (Aliverti 2019, 15-17).

Subject heading languages and thesauri also need to grapple with the issue that what is or is not viewed as being the same topic differs between language or cultural groups, even when the formal structures of the schemes are compatible. Linking pre-existing subject schemes devised according to different structures may best be described as a mapping process. When subject heading mappings have been carefully curated by bilingual cataloguers and the subject heading languages are compatible in structure, the results can be very good. One such project is the European project MACS which linked subject heading authority files in English, French, German, and Italian, where the high level of expertise of the participants avoided erroneous links that could have been caused by false cognates (Landry 2008, 219-220).

The French-language subject heading system used in Canada, the Répertoire de vedettes-matière de l'Université Laval (RVM), originated as a translation of the Library of Congress Subject Headings (LCSH) in 1946, and has retained its parallel structure. The RVM team has carefully maintained the mappings of the RVM headings to LCSH as both systems have evolved (Dolbec 2006; Holley 2002).

JLIS.it

Searching by Subject from the User's Point of View

In a catalogue promoted as bilingual, like Sofia, a user may enter a subject search in their dominant language, without considering that subject access for certain resources may only have been provided in one language and that retrieval using terms from only one language could be incomplete. To avoid putting the responsibility on the user to think of the equivalent terms in multiple languages, Sofia integrates some strategies for expansion of the user's search query with other language equivalents, the most powerful source of valid equivalents for French-English being the RVM authority file. In this file the LCSH equivalent headings are recorded in MARC 21 linking entry fields. This allows indexing English-French subject headings in both directions. Using an RVM authority record with fields 150 and 750 as shown below, a user's search query Musées can be looked up in the RVM authority file, linked to the LCSH form in the linking field, translated to Museums and the query can be expanded to search Musées OR Museums. Using exactly the same fields in the same RVM authority record, a user query for Museums can be looked up in the LCSH linking fields, matched to the RVM accepted form Musées found in the 150 field, and the user's search expanded to search Museums OR Musées.

```
150        __ |a Musées
750        _0 |a Museums
```

If that fails, possibly the user's term does not match an accepted or variant form in the authority file, then a service like Google translate can be called to attempt to provide an equivalent term that can be used in an expanded search. This makes sense for topical subject searching, but not for names or titles, where the best equivalents are to be found in the name authority file.

A pitfall is when a single term in one subject heading language matches multiple terms in the other. This does happen because, as was noted, concepts do not always map cleanly between languages. For query expansion, the system can include all the terms found in the target language in the search. This ensures recall but possibly sacrifices some precision.

Expansion hinges on the accurate identification of the query language, which may not be easy, particularly since the language of the search query may not match the language of the user interface the user is currently working in. The user's search query may be too short to have the language identified, or the string may be ambiguous. For example, information is spelled the same in English and French, and the string "main" has a different meaning depending whether it is interpreted as French (hand) or English (the primary thing).

Expansion intervenes post-cataloguing at the point of the user's search. Another route is to ensure that subject headings in both languages are assigned to bibliographic records, so that all relevant resources will be retrieved whichever language the user searches with. When the records are supplied by different cataloguing agencies depending on the language of cataloguing, completing the subject heading assignment in the other language would require system assistance, either by enriching records in batch or by assisting the cataloguer in finding language-equivalent subjects. The advantage to adding only cataloguer-curated equivalents is mainly for those multiple equivalents. The cataloguer can pick only the one(s) that actually pertain to the resource. All these strategies can be combined and fine-tuned to balance recall with precision, within the practical constraints of cost and time available.

JLIS.it

Concluding Thoughts

In this highly incomplete reflection, I feel that I have presented more issues than answers. Pragmatic approaches that take cost-effectiveness and scalability into account are needed, and that draw the maximum benefit from existing data. A robust approach will need to combine several strategies, compensating for missing metadata by gracefully falling through to alternative mechanisms. There is still much to think about on the road to establishing some best practices for bilingual or multilingual catalogues. I consider that the goal is worth the attempt.

As a final perspective, remember Ranganathan's fourth law of library science: *Save the time of the user.* The system should be doing the work of retrieval, not the user. Even across multiple languages of cataloguing.

JLIS.it

References

Aliverti, Christian. 2019. "Babylonian confusion of languages regardless of standardization? Multilingualism and cataloguing". Presentation slides from 21 August 2019 IFLA WLIC RDA satellite conference, Thessaloniki, Greece. Accessed April 12, 2021. http://www.rda-rsc.org/sites/all/files/aliverti.pdf

Anderson, Dorothy. 1974. *Universal Bibliographic Control: A long term policy – a plan for action.* Pullach/Munich: Verlag Dokumentation.

Cohen Ahava. 2020. "Luck is What Happens When Preparation Meets Opportunity: Building Israel's Multilingual, Multiscript Authority Database". *Cataloging & Classification Quarterly* 58(7): 632–650. DOI: 10.1080/01639374.2020

Dolbec, Denise. 2006. "Le répertoire de vedettes-matière: outil du XXIe siècle." *Documentation et bibliothèques* 52(2): 99–108. DOI: 10.7202/1030013ar

Holley, Robert P. 2002. "The Répertoire de vedettes-matière de l'Université Laval Library, 1946–92: Francophone Subject Access in North America and Europe." *Library Resources & Technical Services* 46(4): 138–149. DOI: 10.5860/lrts.46n4.138

Landry, Patrice. 2009. "La recherche par sujet multilingue dans les catalogues de bibliothèques: la solution MACS." In *Francophonies et bibliothèques: actes du premier congrès de l'Association internationale francophone des bibliothécaires et documentalistes et satellite IFLA, Montréal, 3-6 août 2008,* sous la direction de Dominique Gazo et Réjean Savard, 215–224. Montréal: AIFBD.

Lehtinen, Riitta, and Genevieve Clavel-Merrin. 1998. "Multilingual and multi-character set data in library systems and networks: Experiences and perspectives from Switzerland and Finland." In *Multi-script, multilingual, multi-character issues for the online environment: Proceedings of a Workshop sponsored by the IFLA Section of Cataloguing, Istanbul, Turkey, August 24, 1995*, edited by John D. Byrum, Jr. and Olivia Madison, 67–91. München: K.G. Saur. DOI: 10.1515/9783110948745.67

Park, Jung-ran. 2007. "Cross-Lingual Name and Subject Access: Mechanisms and Challenges." *Library Resources & Technical Services* 51(3): 180–189. DOI: 10.5860/lrts.51n3.180

Riva, Pat. 2020. "Multilingualism in information retrieval systems: the next challenge." In *Mirna Willer: Festschrift*, edited by Tinka Katić and Nives Tomašević, 134-151. Zadar: MorePress.

Tillett, Barbara B. 2008. *A Review of the Feasibility of an International Standard Authority Data Number (ISADN).* Accessed April 12, 2021. https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/pubs/franar-numbering-paper.pdf

Willer, Mirna, and Gordon Dunsire. 2013. *Bibliographic Information Organization in the Semantic Web.* Oxford: Chandos.

JLIS.it

Rethinking bibliographic control in the light of IFLA LRM entities: the ongoing process at the National library of France

Françoise Leresche[(a)]

a) Bibliothèque nationale de France

Contact: Françoise Leresche, francoise.leresche@bnf.fr
Received: 25 August 2021; **Accepted:** 12 September 2021; **First Published:** 15 January 2022

ABSTRACT

When IFLA defined the concept of Universal Bibliographic Control (UBC) during the 1960s, the objective was to describe all resources published worldwide and split this task internationally by developing tools (such as the ISBD and UNIMARC) for the exchange of descriptive metadata. Today libraries are aiming to build web-oriented catalogues, based on the IFLA LRM model: when the ISBD "resource" is split into the WEMI entities, it seems necessary to adopt a new approach toward UBC and to define new criteria.

The BnF has initiated this process. This paper presents which criteria engage BnF's responsibility as a provider of reference metadata identifying an instance of a WEMI entity or an agent. It also presents the quality approach developed by the cataloguing staff in order to reach its objectives and answer the various needs of the metadata users, in a context where the diversity of metadata sources is modifying traditional cataloguing methods. It also investigates the consequences implied by the various stages of the implementation of IFLA LRM by libraries on the exchange of metadata, and concludes with a commitment to maintain the distribution of reusable metadata for all libraries during a period still to be defined.

KEYWORDS

Universal Bibliographic Control; National bibliographic agencies; IFLA LRM model; International cooperation; Digital resources; Quality policy for metadata.

JLIS.it

Lorsque le concept du CBU a été défini par l'IFLA dans les années 1960, il s'agissait d'assurer un recensement le plus exhaustif possible des **publications** au niveau international et de permettre un partage du travail en garantissant les conditions pour l'échange des métadonnées (règles internationales de description des différents types de ressources (ISBD), format international d'échange (UNIMARC)).

Avec le développement des modèles de l'information bibliographique (FR.. et aujourd'hui IFLA-LRM) et la volonté de construire des catalogues « du 21e siècle » orientés vers le web, implémentant à cette fin le modèle LRM et l'éclatement de la « ressource », telle que définie par l'ISBD, en quatre entités WEMI, une nouvelle approche du CBU est nécessaire : si le principe et les objectifs globaux demeurent, comment les atteindre dans le contexte actuel ? Quel domaine d'application en termes d'entités ? Quel rôle et quelle responsabilité des agences bibliographiques nationales sur les instances de ces entités ?

L'expérience du contrôle bibliographique appliqué aux agents

Le développement des fichiers d'autorité, en particulier pour contrôler les points d'accès autorisés représentant les agents (personnes, collectivités, familles) exerçant une responsabilité quelconque par rapport aux ressources décrites a déjà été l'occasion de réfléchir au niveau d'engagement qu'une agence bibliographique nationale peut avoir sur les métadonnées d'identification d'un agent présent dans son catalogue. La réponse couramment adoptée est d'assurer des métadonnées d'identification complètes, faisant référence au niveau international, pour les agents « nationaux » ou considérés comme tels. La nationalité associée à un agent est un attribut qui a été défini et utilisé très tôt dans les fichiers d'autorité français, mais il s'est heurté à une certaine incompréhension au niveau international, la notion de nationalité pouvant varier d'un pays ou d'une culture à l'autre. C'est aujourd'hui la notion plus vague de « pays associé à un agent » qui prévaut au niveau international ; elle est amplement suffisante quand il s'agit de définir les responsabilités en matière de CBU et de dire qu'une agence bibliographique nationale est responsable de l'établissement des métadonnées de référence pour les agents associés au pays dont elle relève.

Quels critères pour le CBU en ce qui concerne les œuvres et les expressions ?

Il semble naturel d'étendre la même logique aux instances des entités présentes dans toute ressource bibliographique au sens de l'ISBD, notamment aux œuvres et aux expressions matérialisées dans les manifestations auxquelles la définition originelle du CBU continue de s'appliquer.

Que signifie exactement cette nouvelle approche et quelles sont ses implications pratiques ?

Dans le cas d'une manifestation publiée en France matérialisant une œuvre d'un auteur français et son expression originale, peu de changements en réalité par rapport à l'approche actuelle : l'identification de référence de la manifestation, mais aussi de l'œuvre et de son expression représentative relève de la responsabilité de l'agence bibliographique nationale française, en l'occurrence de la BnF.
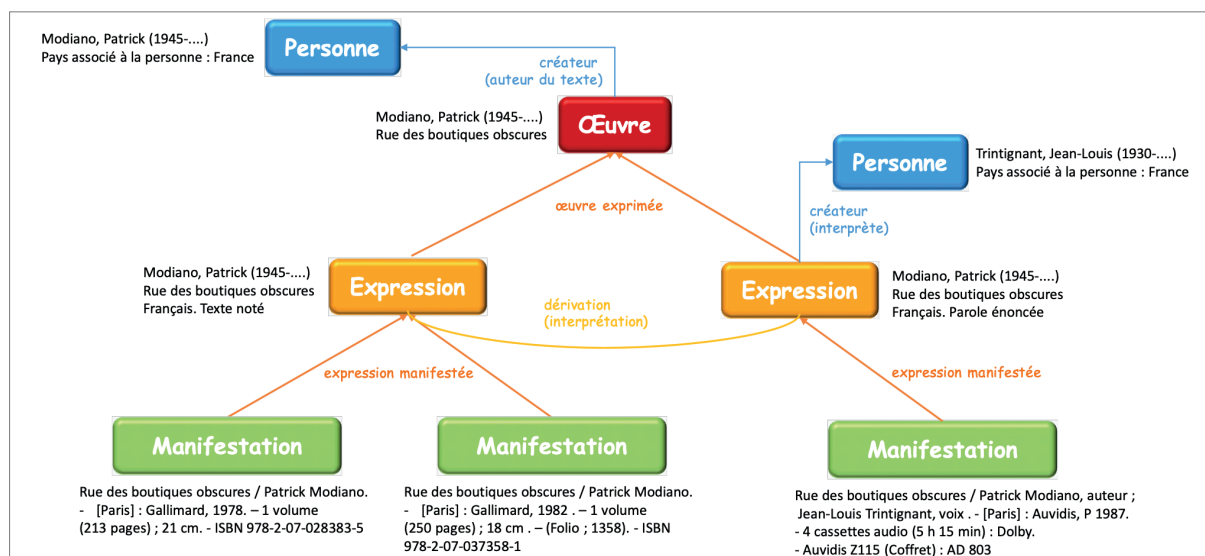
Fig. 1 : Manifestations publiées en France matérialisant l'expression représentative d'une œuvre créée par un auteur français, ainsi qu'une expression dérivée (lecture) de celle-ci : la BnF a la responsabilité d'identifier toutes les instances d'entités présentes dans ce schéma

En revanche, dans le cas d'une manifestation publiée en France contenant une traduction en français d'une œuvre étrangère, la responsabilité de la BnF dans l'identification de référence au niveau international ne s'applique qu'à la manifestation et à l'expression correspondant à la traduction française. L'agence bibliographique nationale française n'a pas de responsabilité particulière en ce qui concerne l'identification d'une œuvre étrangère et peut se limiter à ses besoins fonctionnels.
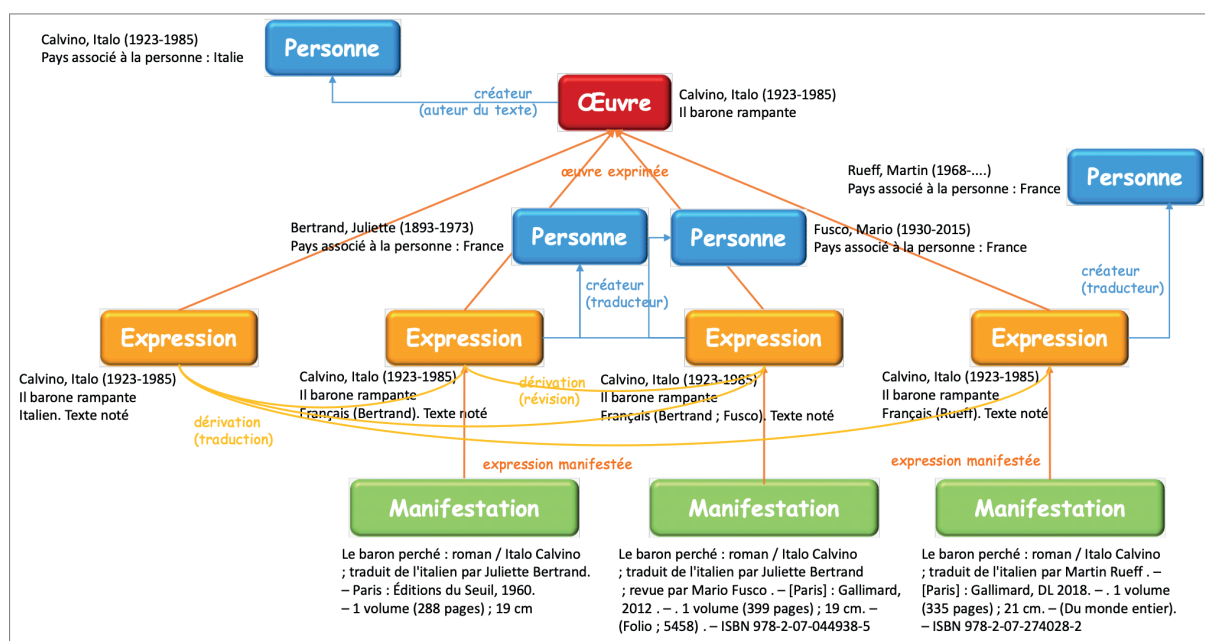

Fig. 2 : Manifestations publiées en France matérialisant des traductions d'une œuvre étrangère, créée par un auteur italien : la BnF a uniquement la responsabilité d'identifier les manifestations publiées en France et les expressions (traductions françaises) qu'elles matérialisent, ainsi que leurs créateurs (traducteurs)

JLIS.it

Dans cette délimitation des responsabilités respectives des agences bibliographiques, si le critère de la langue s'impose spontanément pour les expressions, il n'est pas suffisant. On constate donc la prise en compte croissante de la notion de « pays associé à un agent », puisque c'est le plus souvent le critère le plus objectif dont on dispose pour définir en conséquence le pays associé à une œuvre, mais aussi à une expression. Cela devient donc un critère essentiel, plus important que le pays de publication de la manifestation qui n'est pertinent que pour cette seule entité.

Cette extension semble facile à appliquer tant que l'on s'en tient à un cadre traditionnel : catalogage document en main (identification bibliographique élaborée *ex-nihilo* ou exploitant des métadonnées fournies par les éditeurs) par des catalogueurs pour une production imprimée (texte, musique notée, cartes).

Elle n'est pas aussi aisée à transposer aux ressources audiovisuelles qui soulèvent d'autres questions du fait de leur circuit commercial qui ignore très largement la notion de publication au sens traditionnel : ce qui est pertinent (pour l'image animée comme pour le son) c'est l'étape de la production et celle de la diffusion. En France, le dépôt légal recouvre les ressources diffusées en France, ce qui est extrêmement vaste en ce qui concerne les enregistrements sonores qui peuvent être produits dans le monde entier. Retenir l'échelon de la production ne semble pas non plus pertinent, car dans le domaine de la musique enregistrée la plupart des producteurs sont des sociétés multinationales ou européennes ; quant au cinéma, les coproductions associant plusieurs pays se multiplient.

Les critères qui semblaient simples pour les manifestations imprimées ne s'avèrent pas ou peu pertinents ; il convient donc d'en définir d'autres, en s'appuyant à nouveau sur les critères retenus pour les agents créateurs des œuvres et expressions matérialisées dans les manifestations diffusées en France. C'est une piste qui est envisagée pour le traitement du dépôt légal des ressources audiovisuelles.

Les orientations retenues aujourd'hui par la BnF pour considérer qu'une œuvre ou une expression relève de sa responsabilité d'agence bibliographique nationale au regard du CBU prennent en compte les critères suivants :

- Lieu de publication ou de diffusion de la première manifestation matérialisant l'expression représentative de l'œuvre ;
- Langue de l'expression représentative de l'œuvre ;
- Nationalité des créateurs si le critère de langue ne s'applique pas (image fixe, musique) ou en complément de celui-ci (littérature francophone).

Le défi posé par la multiplication des ressources numériques

Aujourd'hui la diffusion des ressources passe largement par le numérique, de manière massive dans le domaine audiovisuel (enregistrements sonores, films et séries), mais aussi pour les ressources continues et dans une moindre mesure en France pour les ressources dont le circuit traditionnel de publication/diffusion demeure fort (livres, partitions, cartes, etc.). Leur entrée en masse dans les collections s'accompagnent de métadonnées produites en amont, par des acteurs commerciaux dont les objectifs et les pratiques de signalement ne sont pas les mêmes que ceux des bibliothèques.

JLIS.it

À cet égard, les métadonnées associées aux ressources dématérialisées, fournies par des opérateurs commerciaux (les agrégateurs dans le cas du dépôt légal des enregistrements sonores), se caractérisent par :

- leur **hétérogénéité** : selon leur source, les données descriptives peuvent varier en complétude et en structuration, depuis des descriptions minimales, extrêmement pauvres et peu structurées, jusqu'à d'autres présentant une grande finesse de détail (musiciens participant à un ensemble, par exemple). La désambiguïsation des noms cités, en particulier des agents, est loin d'y être une préoccupation largement partagée.

- leur **granularité** : la manifestation comme unité matérielle et intellectuelle fédérant plusieurs contenus, comme c'est majoritairement le cas pour les enregistrements sonores où les agrégats sont la règle, tend à disparaître au profit de l'individualisation de chaque plage, avec un recensement très riche des divers intervenants (créateurs, interprètes, responsabilités techniques et commerciales) à un niveau jamais pratiqué par les bibliothèques ; en revanche, il revient aux bibliothèques de « reconstituer » l'agrégat correspondant à l'album, c'est-à-dire à la manifestation publiée dont il existe souvent un ou plusieurs équivalents sur support.

  Une situation similaire se pose pour les publications en série en ligne où les deux niveaux importants pour les fournisseurs sont le titre d'une part, les articles d'autre part. Le niveau de la livraison (fascicule ou volume) publié à périodicité régulière perd de sa pertinence dans l'univers numérique.

- leur **abondance** : face à l'afflux massif de métadonnées exogènes, il devient impossible d'envisager de les soumettre toutes à un processus de relecture/validation/amélioration par des catalogueurs. Il faut admettre que certaines ne seront pas retravaillées et ne feront pas l'objet d'un processus d'amélioration de la qualité autre que des traitements de masse automatisés, le cas échéant.

Définir une politique de qualité : un outil au service des objectifs du CBU

La BnF s'est dotée depuis de nombreuses années d'une politique de catalogage prenant en compte son rôle d'agence bibliographique nationale chargée d'établir la bibliographie nationale française, politique qu'elle actualise régulièrement pour suivre les évolutions de l'édition comme du contexte bibliographique et technique.

En 2017, elle a pris la décision de transformer son catalogue pour implémenter réellement le modèle IFLA-LRM et permettre la production de métadonnées structurées selon les entités LRM, à commencer par les quatre entités WEMI : le projet NOEMI vise à la création d'un nouvel outil de catalogage permettant de décrire et de lier entre elles les entités LRM. Il s'articule avec le projet national du FNE (Fichier national d'entités), dont l'objectif est de mutualiser la production et la diffusion des données d'identification produites par les bibliothèques françaises, en premier lieu la BnF et le réseau de l'ABES, pour les entités traditionnellement décrites dans des fichiers d'autorité : agents (personnes, familles, collectivités), lieux, concepts gérés dans des listes d'autorité matière, mais aussi œuvres et, à terme, expressions.

En parallèle de ces chantiers, la BnF a engagé une réflexion en vue de définir une politique de

JLIS.it

qualité des métadonnées[1], en s'appuyant sur la modélisation LRM et les tâches utilisateurs définies dans le modèle. Implémenter le modèle LRM (entités et relations) doit permettre d'assurer aux utilisateurs finaux des données de qualité répondant à leurs divers besoins. L'évaluation de la qualité des métadonnées présentes dans le catalogue, qu'elles soient directement produites par les catalogueurs de la BnF ou qu'elles proviennent de réservoirs externes, s'articule autour de différents aspects :

- une **approche par entités** : les instances des entités sont considérées pour elles-mêmes, indépendamment du contexte de catalogage (identifier telle œuvre ou tel agent quels que soient le support ou le type de médiation utilisés dans les manifestations – ce qui permet de se dégager des biais induits par les filières d'entrée du dépôt légal). Cette approche est au cœur du projet du FNE et de la démarche de catalogage partagé qu'il promeut. Elle conduit à doter chaque instance d'entité d'une indication du niveau de qualité qui lui est propre et qui peut différer de celui d'une autre instance qui lui est liée : la qualité est évaluée avec une granularité beaucoup plus fine qu'actuellement où c'est la notice bibliographique dans son ensemble qui se voit affecter un niveau de qualité, souvent lié à la filière de catalogage qui l'a produite (bibliographie nationale française, acquisitions) ;
- la définition de **niveaux différenciés de qualité**, conçus comme des cercles concentriques de qualité, prenant en compte :
  - la *responsabilité au regard du CBU* : identification complète de référence des instances d'entités relevant des critères retenus pour définir une responsabilité d'agence bibliographique nationale, niveaux de qualité moins exigeants et variés pour les autres ;
  - la *capacité à répondre aux tâches utilisateurs* définies dans le modèle IFLA-LRM : construction des points d'accès (points d'accès autorisés et variantes) donnant accès aux instances décrites, identification et enregistrement des relations entre instances (relations fondamentales entre WEMI, relations de responsabilité entre agents et WEMI, relations entre œuvres, entre expressions, entre manifestations), méthodes d'enregistrement de ces relations (note, point d'accès autorisé structuré, identifiant pérenne) ;
  - la *traçabilité des données* en visant, dans la mesure du possible, une granularité au niveau de la donnée : indication de l'origine des métadonnées, des ajouts venant de sources externes (résumés fournis par les éditeurs, par exemple), mais aussi des traitements (manuels ou automatisés) faits sur les données pour en améliorer la qualité, ces traitements portant essentiellement sur les données rétrospectives. Ces informations permettent de juger des métadonnées en fonction des usages de chacun (et des critères de qualité personnalisés associés à ces usages).

Le choix d'implémenter le modèle IFLA-LRM dans le catalogue de la BnF est considéré comme un gain en efficience du fait de la factorisation de certaines informations au niveau de l'œuvre (indexation matière, relations entre agents et œuvre) ou de l'expression (dépouillement des agrégats, relations entre agents et expressions), particulièrement utile dans le cas de manifestations multiples (simultanées ou successives), mais aussi comme un gage de qualité en termes de complétude et de cohérence des données au sein du catalogue.

---

[1] La politique de qualité des métadonnées s'articule avec la politique des identifiants (voir la communication de Vincent Boulet *How to build an «Identifiers' policy»: the BnF use case*, publiée dans ce numéro de JLIS.it).

JLIS.it

La référence au modèle IFLA-LRM est aussi un gage d'interopérabilité avec les autres bibliothèques, au-delà de choix d'implémentation différents (raccourcis, etc.), mais aussi avec d'autres communautés professionnelles dans le domaine de la culture, notamment les archives et les musées.

Assurer la transition

Le CBU repose sur le principe du partage du travail de recension et de description, avec pour corollaire l'échange des données entre les pays et les agences bibliographiques. Passer d'une logique de description bibliographique de ressources, telles que définies par l'ISBD, à une structure par entités LRM liées entre elles (structure relationnelle) pose un problème pour l'échange, du fait de la diversité des situations parmi les agences bibliographiques. Si aujourd'hui les catalogues articulés autour de notices bibliographiques et de notices d'autorité liées (ou non) sont majoritaires, le passage vers des bases de données relationnelles structurées selon les entités LRM va se faire progressivement, mais à des rythmes différents et selon des modalités et des formats variés. La continuité des échanges entre agences bibliographiques, ayant fait des choix d'implémentation différents selon des calendriers qui leur sont propres va nécessiter d'assurer une période de transition où les données produites sous forme LRMisées devront être converties pour fournir des notices bibliographiques conformes à l'ISBD et des notices d'autorité liées, selon les modalités de diffusion actuelles.

Les deux agences bibliographiques françaises, l'ABES et la BnF, s'y sont engagées auprès des bibliothèques françaises dans le cadre du programme national de la Transition bibliographique. Les bibliothèques étrangères pourront naturellement en profiter, mais cette double fourniture des données bibliographiques aura une durée limitée dans le temps, en fonction de l'évolution des catalogues des bibliothèques françaises vers la nouvelle structure LRMisée.

En parallèle, les deux agences travaillent ensemble au sein du CfU à faire évoluer le format UNIMARC (format bibliographique et format d'autorité) pour lui permettre de rendre compte des entités LRM et de leurs relations. L'objectif est que, quelle que soit la structure de leur catalogue, les bibliothèques puissent continuer à disposer d'un format international d'échange, riche et précis, pour échanger les données bibliographiques qu'elles produisent et/ou réutilisent, selon les objectifs du CBU qui demeurent par-delà des changements technologiques qui ont transformé le contexte des catalogues de bibliothèques.

Remerciements

JLIS.it

Bibliographie

Anderson, Dorothy. 1974. *Universal Bibliographic Control: a Long Term Policy, a Plan for Action.* Pullach bei München: Verlag Dokumentation.

Bibliothèque nationale de France. 2018. *Politique de qualité des données.* Consulté le 15 juillet 2021. Disponible en ligne: https://www.bnf.fr/fr/politique-de-qualite-des-donnees.

IFLA. 2012. *Professional Statement on Bibliographic Universal Control.* Consulté le 15 juillet 2021. Disponible en ligne: http://www.ifla.org/files/assets/bibliography/Documents/ifla-professional-statement-on-ubc-en.pdf.

IFLA. 1961. *« Principes de Paris » adoptés par la Conférence internationale sur les Principes de catalogage, Paris, Octobre 1961.* Consulté le 15 juillet 2021. Disponible en ligne: https://www.ifla.org/files/assets/cataloguing/IMEICC/IMEICC1/statement_principles_paris_1961.pdf. Traduction française disponible en ligne: https://www.ifla.org/files/assets/cataloguing/IMEICC/IMEICC1/statement_principles_paris_1961-fr.pdf.

IFLA Cataloguing Section and IFLA Meetings of Experts on an International Cataloguing Code. 2016. *Statement of International Cataloguing Principles: ICP.* 2016 edition with minor revisions, 2017. Consulté le 15 juillet 2021. Disponible en ligne: https://www.ifla.org/files/assets/cataloguing/icp/icp_2016-en.pdf.

IFLA FRBR Review Group. Consolidation Editorial Group. 2017. *IFLA Library Reference Model: a conceptual model for bibliographic information.* Disponible en ligne: https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf. Consulté le 15 juillet 2021.

Illien, Gildas et Bourdon, Françoise. 2014. *À la recherche du temps perdu, retour vers le futur: CBU 2.0.* Communication présentée à: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge, Session 86 - Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Programme. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. Consulté le 15 juillet 2021. Disponible en ligne: http://library.ifla.org/956/1/086-illien-fr.pdf. Traduction anglaise disponible en ligne: http://library.ifla.org/956/7/086-illien-en.pdf.

JLIS.it

The future of bibliographic services in light of new concepts of authority control

Michele Casalini[(a)]

a) Casalini Libri, http://orcid.org/0000-0003-4643-8895

ABSTRACT

Over the last three decades, a number of major changes in the field of cataloguing have led to the definition of new forms of authority control. The introduction of FRBR and of IFLA LRM have been followed by continuing studies, including, more recently, the implementation of linked data in library catalogues, as well as improvements to data models in order to ensure the broadest possible interoperability among systems. A new approach to authority control and its connected services can be based on the combination of manual and automatic processes of data validation and enrichment, together with the use of knowledge bases as authoritative sources. This will also grant wider data interoperability, opening up a new level of cooperation among the international institutions and organisations concerned with the dissemination of knowledge.

KEYWORDS

Authority control; Bibliographic metadata; Cataloguing ecosystem; Linked Open Data (LOD).

JLIS.it

New era, new needs

The surprising technological innovations and significant changes in the field of cataloguing have opened the doors to new horizons that see machines play a proactive and effective role in the decoding and sharing of bibliographic metadata. It is thanks to these new advances in technology that it is possible to overcome linguistic barriers and venture beyond purely bibliographic fields.

In the new, digital, era, the fast growing quantity of – sometimes perishable – data, requires those who operate in the cultural heritage sector to carry out a task of fundamental importance: to react to the need for an authority control that "guarantees" the homogeneity, stability and formal quality of access entries as an integral operation within the cataloguing ecosystem. It is a technique influenced by the technology of its time as well as by the standards and cataloguing conventions in use in the contexts of linguistic, cultural and disciplinary specializations.

The concept of traditional authority records is also evolving, in order to comply with the new open philosophy of data sharing and reuse. The transition from the concept of record to that of entity, in the context of the semantic web has forced a rethinking not only of data, but also of the organization and management of authority control itself. Previous discussion on whether authority control should be based centrally or locally will be subject to transformation, as the focus shifts from a rigid conception to a more flexible notion of entity identification and relationships between entities. The direction in which this field is advancing has already been partially outlined in the enlightening profile within the international conference proceedings of the Authority Control in Organizing and Accessing Information, held in Florence in 2003.

In combination with the technological developments that support this cause, since then it has been necessary to radically rethink the conceptual models of data interpretation. The transition from the Functional Requirements for Bibliographic Records (FRBR) to the IFLA Library Reference Model (IFLA LRM) has propelled the international community towards a new modeling of bibliographic levels, linked together by primary relationships and accompanied by further relationships with entities and properties.

The Bibliographic Control function continues to be valid today but shifts the focus to a global level, supporting growing international cooperation, which is facilitated now by the interoperability of the data models. The contribution of a heterogeneous group of organizations concerned with the dissemination of knowledge also promotes cooperative authority control, with collaboration and mutual assistance among actors of various kinds; by comparing and integrating their data with those of others, the information they convey will be more complete and more reliable.

Organizations such as libraries, archives, museums, but also publishers and providers will engage with each other in the generation of new data and the discovery of new resources, crossing the boundaries of specific domains to create data enrichment opportunities that would previously have been unthinkable. The theme of facilitating the sharing of authoritative sources through persistent and reconciled resources for the benefit of a more precise and wider discoverability was also addressed from 2016 to 2018 by the Institute of Museum and Library Services (IMLS) funded National Strategy for Sharable Local Name Authority National Forum (SLNA-NF).

To implement the interoperability of metadata, it has become necessary to create a new conformation and structure, so that each entity can be identifiable by a single and unambiguous name or code that is used by all agencies creating bibliographic metadata: the Uniform Resource Identifier

JLIS.it

(URI) avoids the ambiguity of using natural language. Data structured according to the Resource Description Framework (RDF) data model, in contrast to the traditional record-based approach, focuses on individual metadata declarations represented by triples of data in the subject-predicate-object form.

These triples can become quads, containing the provenance information necessary to take advantage of data enriched through authoritative sources, while maintaining local preferences for the labeling and display of data through customizable application profiles.

Statements can be combined and matched from many different sources to link different standards and models as well, such as Resource Description and Access (RDA) and the Bibliographic Framework Initiative (BIBFRAME). The schemas expressed in the RDF linked data structure allow other communities to reuse the data in their own environments.

New models into practice: the Share Family

Following the path that was initiated, developed and progressively applied by the Library of Congress with BIBFRAME, encouraged by the vision of the Linked Data for Production (LD4P) projects promoted by Stanford University, and in light of the extensive and exciting possibilities offered by new technologies and data models, in 2016 a community-driven initiative, the Share Virtual Discovery Environment (Share-VDE or SVDE), emerged, with the aim of putting the new developments into practice and applying them to an entity based discovery environment for the benefit of libraries and their users.

As one of the founding organizations of the Share-VDE initiative, and in its role as a bibliographic agency, Casalini Libri has been, and continues to be active in testing linked data technologies for libraries together with its technological arm and sister company @Cult.

Building on the experience of all involved parties and drawing from it, one of the aims of SVDE has been to develop innovative approaches for the authority control of bibliographic records and for the creation and improvement of authority control procedures, providing new authority services to libraries and supporting their transition to linked data.

The starting point for this evolution, like the initiative itself, stems from the real and emerging needs of the library community, more specifically the need for libraries to receive constantly updated information on their bibliographic and authority records from authoritative sources, both in MAchine Readable Cataloguing (MARC) format and in the BIBFRAME linked data structure. The services designed and the underlying technological infrastructure are the result of the development of new Linked Open Data (LOD) technologies influenced by the direct input of the various Share Family collaborative environments involving national and research libraries. These processes facilitated the experimentation in the creation and handling of linked data entities, but also provide direct interaction with operational library systems that will coexist for a long time in both MARC and RDF.

The overall goals include the enrichment of MARC records with identifiers from external sources (e.g. ISNI, VIAF); the reconciliation and clusterization of entities and the publication of the Cluster Knowledge Base (CKB); the conversion from MARC to RDF using the BIBFRAME vocabulary together with other ontologies; batch/automated authority services, data updating and data

JLIS.it

dissemination procedures; a manual entity management tool (J.Cricket); the publication of data on an entity-oriented user interface (www.svde.org).

An active role in determining directions and priorities is played by the Share-VDE Advisory Council and by the various Working Groups, one of which is dedicated specifically to the Authority/ Identifier Management Services (AIMS).

Flexibility in handling and in profiling the integration of data from external sources is a crucial aspect for the processes involved, as each institution may have a different list of priorities.
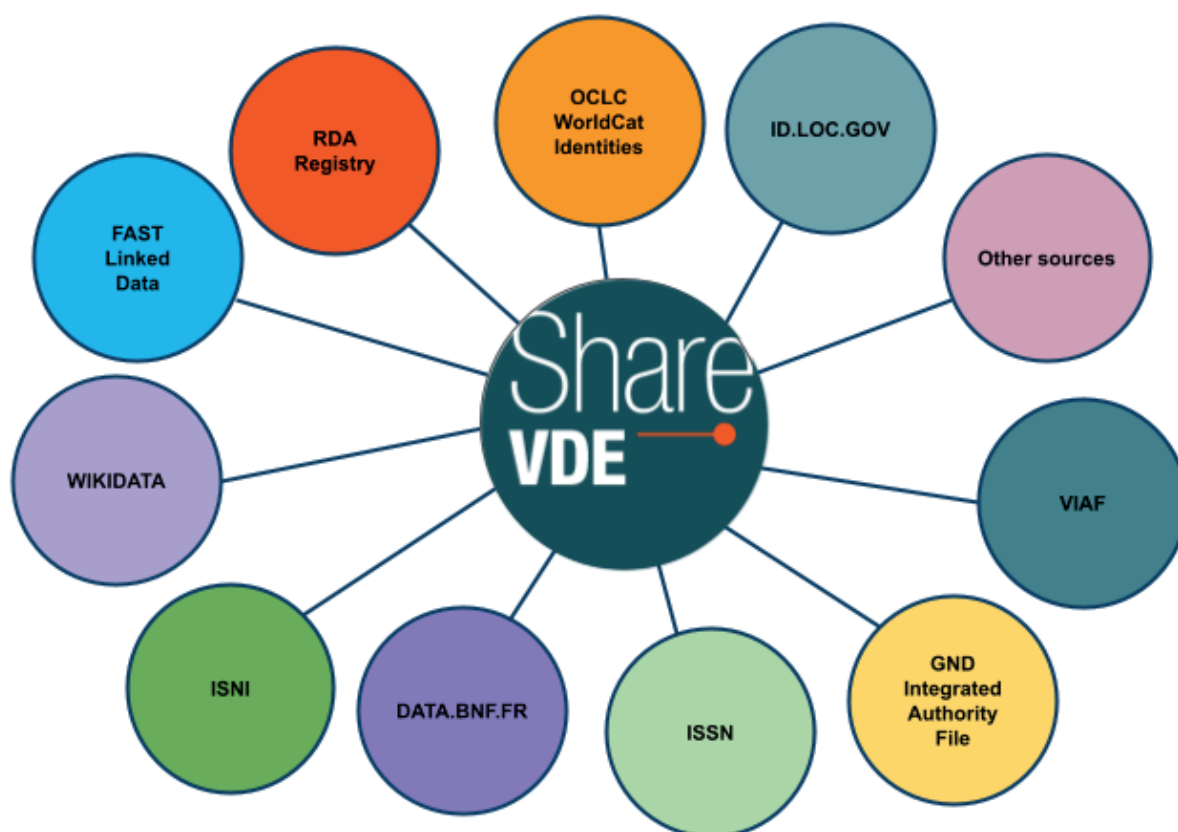
Fig. 1. Integration of data from external sources

Wikidata is an example of interaction that allows for sources to be searched and for SVDE data to be enriched with Wikidata entity information – and vice versa – as SVDE has a property in Wikidata for the author ID.

From a technological viewpoint, Application Programming Interfaces (APIs) architecture simplifies interconnections, reusability, sustainability and scalability, opening the window to an open world.

JLIS.it

The challenge of data models interoperability

The challenge of data interoperability among systems, which is indispensable in order to bring into practice implementations at a wide scale, however, requires comparisons among data models and mapping that maintain the granularity of information. With this aim, on June 10th 2020 the SVDE Entity Identification Working Group approved the SVDE Opus class, also a BIBFRAME Work, as the SVDE Work is too.
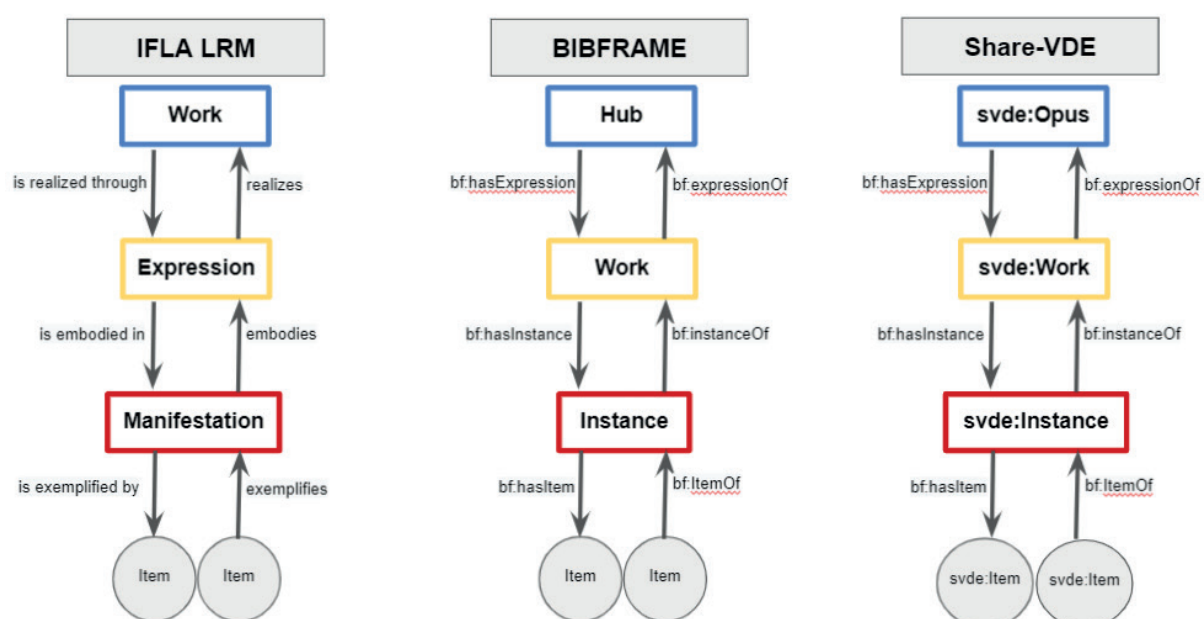

Fig. 2. IFLA LRM, BIBFRAME, Share-VDE Data model comparison. Further details on this structure can be found on https://wiki.svde.org

It is important to highlight that SVDE is trying to practically reconcile an approach to entity modelling that is "North-American oriented" (BIBFRAME) with a "Europe-centric" approach (IFLA LRM). This reconciliation aims to create a flexible crossover between different cataloguing practices, thus allowing it to adapt to different data modelling contexts that cannot be confined in restricted geographic, linguistic, cultural borders. Such trait d'union has been facilitated by the entry in SVDE of European libraries such as the National Library of Norway, the National Library of Finland and the British Library.

We have now mentioned several of the pillars that relate to one another and create the broader ecosystem with the Share-VDE Cluster Knowledge Base, named Sapientia, in the center.
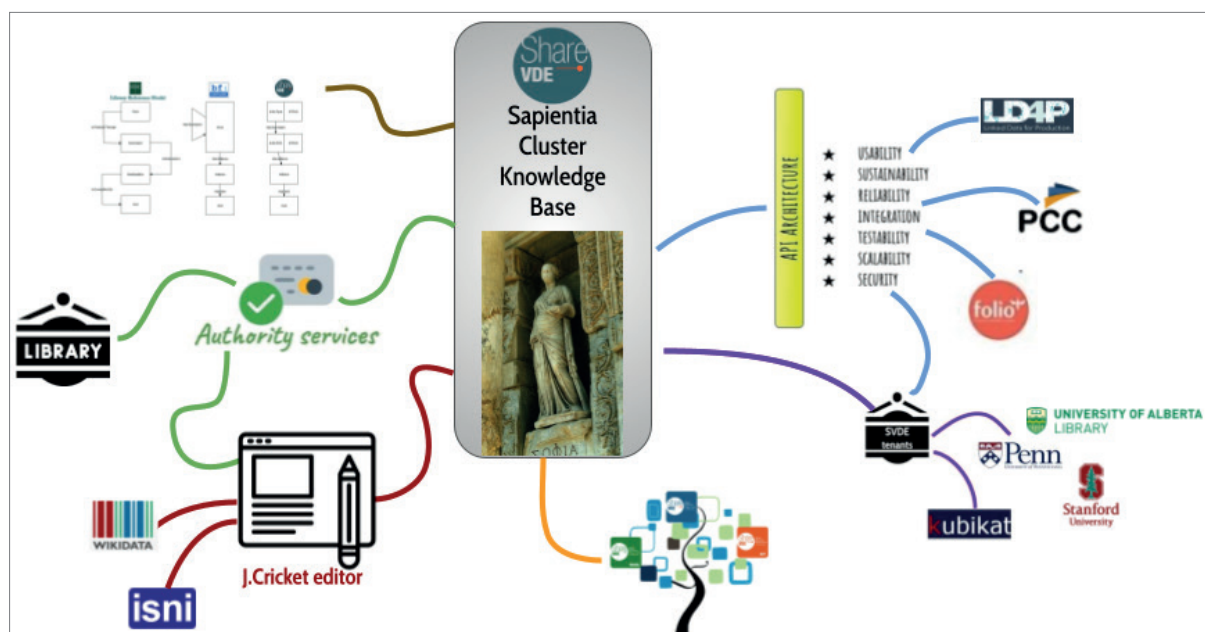
JLIS.it

Fig. 3. Sapientia Cluster Knowledge Base ecosystem

Around it, in a clockwise direction, are the APIs layer (back-end to interact with other environments), the skin and tenant architecture, the authority flows handling both automatic and manual processes, traditional data flows in MARC or entity oriented ones, and factors regarding interoperability with other data models.

The CKB, Sapientia, represents the ambition to build an authoritative knowledge base with the tools to improve it, with 1) mechanical algorithms but also 2) an editor, J.Cricket, which allows a transversal community including data producers (e.g. librarians) to collaborate jointly in order to raise the level of quality of the data offered. The combination between the CKB, produced through automated processes that aggregate data from many different sources, and its editor could represent a sort of ideal union of the Virtual International Authority File (VIAF) world and the Wikidata community; that is, on the one hand the quality and authority proposed by the VIAF model, and on the other hand the open and cross-domain approach of the Wikidata model, carrying its vision of a collaborative tool open to a wider community of users. Therefore Sapientia, with its control and editing tool, J.Cricket, opens up to a vision that combines authority with the ability to interact: openness and control.

Authority processes in the new bibliographic ecosystem

In the new context, highlighted above, the aim of authority control services is to facilitate the control and standardization of bibliographic data. This is achieved through a combination of automatic and manual processes that make it possible for local cataloguing practices to be integrated within a global, participatory dimension.

Automatic authority control operations allow a high level of productivity, while manual operations

JLIS.it

guarantee a higher level of quality: for each context, therefore, the best balance of the two must be identified.

The automated processes are divided according to whether the library is handling a record-based ILS, an RDF-based system, or a hybrid one. These are some of the processes involved. Variables may be the frequency of the dataflows and whether the library also holds the authority file locally.

- The MARC record validator, the MARC corrections for errors and obsolete forms, and the matching/enrichment with profile sources compose the record-based scenario.
- Access point enrichment (including Series and Subjects), matching, import and interaction with the Sapientia Cluster Knowledge Base are necessary for interaction with the RDF-based systems.

In both cases the processes are enabled through Representational State Transfer (RESTful) modules of the LOD Platform, which provide bibliographic, authority and full text search services with entity detection and identification including relator terms capabilities.

The manual processes are divided into two groups: the operational tasks of the original cataloguing processes to validate and enrich metadata elements, and the editing processes to enhance the common Cluster Knowledge Base.

The first set of manual processes can be characterized by the following operations:

- Authority control of the access points of bibliographic records for similar matches and non matches, including the checking, validation and reconciliation of imported URIs.
- Manual enrichment of entity Work and Agents (including Publishers).
- The creation of original authority records; Casalini Libri already does this for the International Standard Name Identifier (ISNI), in compliance with its role as an ISNI Registration Agency (Personal Names, Corporate Names, including publishers, Meeting Names and Uniform Title) and sends reports to ISNI in the case of duplicate records existing for a single entity or of relationships with incorrect titles.

These operations are enabled through the dedicated URI Registration Platform.

The second set of manual processes employs the CKB editor, known as J.Cricket, as the instrument for the direct management of entities represented in RDF. The new application, dedicated to the editing of SVDE community data, is a collaborative tool that not only makes it possible to validate automatic matches that the clustering procedure identifies as uncertain, but also allows library professionals to merge, split or create new clusters autonomously. Conceived as a collaborative editing environment, the application foresees different levels of access and interaction with the data, enabling users to manually create, modify and reconcile clusters of the entities saved in the CKB.

The entities present in Sapientia and managed by J.Cricket are based – conceptually – on the SVDE four labelled entity model (Opus, Work, Instance, Item). The clusters to be modified, automatically and manually, are: Opus, Works and Agents. The next achievement will be to treat the Instance as an entity.

These two examples show how the interaction between J.Cricket and Wikidata IDs is envisioned, from both perspectives.

The scope and capacity of the CKB editor will be extended over time to include the management of authority services for libraries, with quality control procedures for data. With this twofold purpose, not only will J.Cricket facilitate the creation and handling of linked data entities within SVDE, but it will also provide direct interaction with library systems both in MARC and RDF formats.

JLIS.it

Interconnections both with the Sinopia linked data cataloguing module of the LD4P initiative and with the data from the Program for Cooperative Cataloguing (PCC) will be the primary testbed for J.Cricket to prove its ability to act as a pivotal tool between traditional MARC-based cataloguing workflows and innovative linked data processes.

Linked data for the future

The linked data paradigm is laying the groundwork for new level of cooperation among international organizations to create new bridges across the library, archives and museums domain, which serve to increase discoverability for students, scholars and the wider community, to reveal data that would be otherwise remain hidden, to contribute to promoting a culture of openness towards knowledge, and to foster – on the one hand – the preservation of existing knowledge and – on the other – the progress undertaken by younger generations.

Initiatives such as LD4P, Share-VDE and others with each of the institutions involved, the leading role of many national libraries, of cooperative programs such as the PCC, and of other players in the information chain are crucial not only for bringing these developments into practice, but for reaching the critical mass of implementation across cultural heritage collections.

In conclusion, the present challenge for the organizations that have bibliographic control at heart is not only to facilitate libraries in handling constantly updated information on their records or datasets from authoritative sources, but also to improve the level of collaboration between actors of differing nature, thanks to data interoperability, in a future vision of authority control which is more open and cooperative on a global scale.

Acknowledgements

JLIS.it

References

Taylor, Arlene G, and Barbara B. Tillett, eds. 2004. *Authority Control in Organizing and Accessing Information: Definition and International Experience.* New York: The Haworth Information Press. https://doi.org/10.4324/9780203051092

Bibliographic Framework Initiative (BIBFRAME). Accessed June 3, 2021. https://www.loc.gov/bibframe

IFLA Study Group on Functional Requirements for Bibliographic Records. Functional Requirements for Bibliographic Records (FRBR): Final Report. 1998. Munchen: K.G. Saur. https://www.ifla.org/publications/functional-requirements-for-bibliographic-records

IFLA Library Reference Model (LRM). 2017. Den Haag: International Federation of Library Associations and Institutions. https://www.ifla.org/publications/node/11412

International Standard Name Identifier (ISNI). Accessed June 3, 2021. https://isni.org

Linked Data for Production (LD4P). Accessed June 3, 2021. https://wiki.lyrasis.org/display/ld4lGW

MARC 21 Format for Authority data. Accessed June 3, 2021. https://www.loc.gov/marc/authority

Casalini, Michele, Chiat Naun Chew, Chad Cluff, Michelle Durocher, Steven Folsom, Paul Frank, Janifer Gatenby, Jean Godby, Jason Kovari, Nancy Lorimer, Clifford Lynch, Peter Murray, Jeremy Myntti, Anna Neatrour, Cory Nimer, Suzanne Pilsk, Daniel Pitti, Isabel Quintana, Jing Wang, and Simeon Warner. National Strategy for Shareable Local Name Authorities National Forum [SLNA-NF]: White Paper. 2018. https://hdl.handle.net/1813/56343

Program for Cooperative Cataloging (PCC). Accessed June 3, 2021. https://www.loc.gov/aba/pcc/

Resource Description and Access (RDA). Accessed June 3, 2021. https://www.rdatoolkit.org

Share Virtual Discovery Environment (Share-VDE or SVDE). Accessed June 3, 2021. https://wiki.svde.org

International Federation of Library Associations and Institutions. *UNIMARC manual: authorities format.* 2009. 3rd ed. Munchen: K.G. Saur. https://www.ifla.org/publications/ifla-series-on-bibliographic-control-38

Virtual International Authority File (VIAF). Accessed June 3, 2021. https://viaf.org

JLIS.it

New Challenges in Metadata Management between Publishers and Libraries

Pietro Attanasio[a]

a) Associazione Italiana Editori, http://orcid.org/0000-0001-7410-6682

ABSTRACT

Identifiers, bibliographic metadata, thematic category schemes are at the heart of the functioning of the book supply chain. There are international standards for all these elements, which allowed e-commerce to develop in the book trade before any other sector.

The dialogue on metadata management between the book industry and the library community is not always as intensive as desirable. The challenges that the whole book world must cope with today and in the near future pressure us into change. Building on lessons learned from the past, the article focuses on some upcoming challenges, such as big data and artificial intelligence applications, with the aim of identifying fields for a future collaboration.

KEYWORDS

Metadata; Publishing; Big data; Artificial intelligence.

JLIS.it

Introduction

The article focuses on metadata management from the publishing industry point of view, which is slightly different from that of the library community. In the first part I introduce the work made in this field by the Italian publishers association (AIE) and describe our approach in order to identify the reasons why the library approach is different. This is a prerequisite to setting a strategy to bridge the gap between the two.

In the second part I focus on the factors that today are disrupting the traditional context, which are related to the entrance of new players in the book sector and to the impact of big data (vs. metadata) and artificial intelligence.

I conclude that the changes that are occurring call both the book industry and the library community to build a new alliance for a fair and open book data management, starting from some core principles that we share, notwithstanding the differences between commercial purposes and the public sector mission, which will remain.

Publishers approach to book data

The Associazione Italiana Editori (AIE, the Italian publishers association), besides being a trade association representing the Italian publishers' interests at a national and international level, is characterized by a peculiarity which is probably unique: we have a research and development team within the association that is primarily engaged in the fields of book standards and metadata. We develop technologies in these areas, with particular attention – in the last 10 years – to the management of rights metadata, in line with the principle of the Copyright infrastructure launched by the European Council in 2019 and then indicated by the European Commission in relation to the European recovery and resilience plan. The AIE R&D team has been coordinating important European initiatives in the field, such as ARROW – dedicated to the management of rights metadata in digital library initiatives – and the more trade oriented ARDITO.

Linked to this experience, AIE representatives have been and still are in the governance bodies of standard setting organisations such as EDItEUR, ISBN International Agency, IDF (International DOI Foundation), W3C Digital Publishing Business Group, and EDR-Lab (European Digital Reading Lab).

According to our approach, metadata originate from events. Therefore, we place the "event" – rather than the "document" – at the core of our metadata analysis[1]. In this view, metadata start existing before a book is published (or, in general, before any document is produced). The first event to be considered is: "author *A* creates the work *W*", which is relevant even before publishing that work in the form of a document. Such an event originates the need for:

---

[1] This is the ontological difference between the <in*d*ecs> data model and the FRBR. See Rust and Bide (2000), in particular chapter 4.3. "The Commerce View", where the role of the events is described in the terms used in this article. In the FRBR model the *event* is instead one of the "entities that serve as the subjects of intellectual or artistic endeavour". Cf. IFLA (1997).

JLIS.it

a) Uniquely identifying A and W, e.g. with an ISTC[2] and an ISNI;
b) Metadata for describing A and W;
c) A qualifier to identify the relation between A and W: in this case: "A is the author of W".

The second event in the typical life of a literary work is "*A assigns publication rights PR in W to publisher P*", which generates similar metadata needs, i.e. identification and description for the assigned rights and the publisher. Every following event generates needs for new metadata for further editions, translations, transposition for cinema or theatre etc.

More in general, these events can be described as "People make stuff" in the first case, and "People do deals about stuff"[3], in the second case.

Saying that metadata *originate* from events does not mean that metadata are directly *generated* by the events. A common definition of metadata is "An item of metadata is a relationship that someone claims to exist between two entities", which emphasises that there is a level of discretion in making that claim, and thus "the identification of the person making the claim is as significant as the identification of any other entity" (Rust and Bide 2000).

Since metadata are "claims", the objective of the claimer is as important as the nature of the relationships that are described. To understand differences and similarities between the approach to metadata of publishers and that of librarians, it is useful to look at the purposes of the "claimers" in the two cases.

The first purpose in metadata management in the industry is to increase the efficiency in the supply chain. Typically, an important metadata item in our world is the weight of the book, a crucial piece of information to maximize the efficiency of logistics. But the main data items that make a difference between a books-in-print database in a specific country and – for example – the national bibliography in that same country are the book price and its availability (P&A). This little difference (it is a matter of few metadata items) creates a big distance in the management of the two catalogues. P&A data are subject to change over time, which does not happen for other metadata[4], and this implies that a books-in-print database must manage changes in the existing records on a daily basis, whilst the national bibliography is enriched with new titles but the existing records change rarely.

If the need to serve the supply chain determines a big difference, improving the discoverability of books is the main objective that the two communities have in common. Both the industry and libraries need to assist their clients (book buyers or library users) by facilitating as much as possible how they look for and find books. Books-in-print databases and library OPACs shared this pur-

---

[2] The International Standard Text-work Code (ISTC) was the ISO standard to uniquely identify text-based work. Because of very limited use by the industry the standard has been recently withdrawn, though the need for identifying text-works remain. The International Standard Name Identifier (ISNI) is the ISO standard for identifying contributors to creative works and those active in their distribution. See https://isni.org.

[3] Rust and Bide (2000), p. 4. See also Paskin (2006).

[4] Since metadata are "claims" about a relationship, all metadata are not written in stone: a claim may change if there was a mistake or if there is a change in the way claims are expressed in a standard metadata language. In the case of P&A, however, there are continuously new events that originate new relationships and thus the need for new metadata. Prices may change from time to time, and availability changes continuously, both at manifestation and at work level. When dealing with digital library programmes, the metadata element "the work W is out of commerce" is very important and in the EU carries important juridical consequences, after the approval of Directive 790/2019.

JLIS.it

pose since the origins, back in the Seventies. With the advent of the Web this aspect became even more crucial in any service provided to readers. In both communities the awareness on the importance of quality and richness of descriptive metadata grew in last 25 years. The Internet made the role of metadata in search engines crystal clear: to improve discoverability and to provide data to readers to allow them to make informed decisions. In spite of this, there are still differences in one crucial aspect related to discoverability: the subject classification scheme. In particular, in my opinion, the library world did not pay a desirable level of attention to the big effort of the industry to build Thema[5].

The third purpose for metadata is to elaborate statistics about the use of books. In the language I am using, metadata serve to build data about the third kind of events cited in the <in*d*ecs> model, when "People use Works", i.e. when a person buys, or borrows or makes any use or re-use of a book. Statistics are useful to make decisions both for publishers and librarians. The difference, here, is in the perception of the value of a standard vocabulary. Since sales data are produced further down in the supply chain, publishers need standard ways to collect them. Conversely, any library produces data from its users directly, and standardisation is needed only for comparisons with other libraries. This has created more standardisation needs in the trade than in the library world.

The disruption: from metadata to big-data

Metadata, in the traditional meaning understood by publishers and librarians, played an important role in the first phase of the Internet. In mid-Nineties, the book sector was the only sector that had databases containing standard identification and rich description of millions of items, ready to be posted on the Internet, and standard messaging for tele-ordering. This was the reason why e-commerce was developed for selling books before any other good or service. Similarly, library OPACs were the first public service transferred online, in the same years.

The context was disrupted by the (so-called) Web 2.0, i.e. when the Internet started to be characterised by the meta-intermediation of web platforms on one side and user-generated content on the other side[6]. Tracking events of the kind "people-use-stuff" opened a completely different scenario.

Let me start from one specific event:

> (A) Reader *R buys* books *B1*, *B2* and *B3* in bookshop *BS*

Such a simple event generates a number of data:
- The relation "buy" between R and each of the 3 books;
- The relation among the 3 books due to the circumstance that they were bought during the same event;

---

[5] See <https://ns.editeur.org/thema/en>. A short illustration of the origin, purpose and main characteristics of Thema is in Bell and Saynor, 2018.

[6] The evolution between the two phases is well narrated by Foer, 2018. A brilliant - though not rigorous, from a scientific point of view - description of the same evolution is in Lanier, 2011 and 2019 and in many posts of the same author here: www.jaronlanier.com.

- The relation between the 3 books and BS;
- The relation between R and BS.

The two people (the natural person R and the legal person BS) and the 3 manifestations are (or could be) described by metadata, which per se multiply the relationships between the entities. E.g.: if R is 28, a graduate, an Italian citizen, living in Rome, etc.; this creates a relation between all the metadata items of R and the 3 books, and all the metadata associated with each of the 3 books (e.g. all the 3 books are crime novels).

- These metadata may be registered in different sources:
- R may have a BS fidelity card where that information is registered;
- The books' metadata are in a books-in-print database;
- BS metadata are in the database of the Italian bookshops.

Later on, R borrows a book from the public library L, where he/she is registered with another data-set. Then R posts a comment on social media SM about one of those books…

Collecting data of this sort is not new. It is the basis of any statistic on reading, to estimate, for example, how much young, well-educated Italians like reading crime novels. Which was usually done by interviewing a sample of readers.

The disruption lies in the fact that machines are now able to track millions of similar events and the current computing power and memory allow to elaborate all the generated data through powerful algorithms. In principle this allows to collect data about events involving millions of readers that buy or borrow books and post comments etc. All in all, we have billions of data generated by events that machines are able to track.

Combining human intelligence and professional skills with good algorithms, such big data would enable publishers to design outstanding editorial plans and marketing strategies, and librarians to have the perfect collection and reading promotion strategies for their patrons.

Are we still speaking about metadata? If we consider the <in*d*ecs> definition above ("An item of metadata is a relationship that someone claims to exist between two entities") we can easy appreciate the difference: in registering the events here described we have not "someone claiming": it is a matter of machines registering events and extracting data from the events, usually according a pre-defined model[7].

Opportunity or threat?

Machines are able to track any event in our life. Tracking what we read is a very delicate issue, since it involves our thoughts, our lifestyle, our opinions and thus our fundamental rights of freedom of thought and expression. The issue should be treated with all possible care.

In the examples above, R participated in events that produced data which were then controlled by a bookshop, a library and a social media platform (BS, L and SM), each independent from each other. Only R has all the information about the whole picture, and legislation limits the possibility of BS, L and SM to exchange (personal) data about R.

---

[7] Machines may also produce metadata as defined in the <in*d*ecs> model. There is extensive literature about the automatic extraction of metadata (keywords, subject, etc.) from texts. See, for example, the recent Li 2021, useful also for the reference list. In this case there is "someone claiming": it is the machine, with the algorithm or, better, the person who runs the machine for that purpose.

JLIS.it

At the same time, data have an economic value, and determine more and more market power. When R buys all books from one Internet shop, together with many other goods, and posts reviews of the books in the same shop, and uses the cloud services and the platform of the same company for audiobooks, e-books and videos, etc., that single company acquires information and know-how that other competitors can never reach. Data control is a key driver to market power in the digital economy, as is also recognised by the proposal for a Regulation on Contestable and fair markets in the digital sector (known as DMA – Digital Markets Act), which emphasises the presence of "data driven advantages" (Recital 2), the existence of barriers-to-entry generated by data control (Rec. 3), stressing the "potential advantages in terms of accumulation of data, thereby raising barriers to entry" (Rec. 36)[8].

The reasons why data are so relevant in digital markets are well explained by the literature. "The quintessential task of many digital platforms is that of making predictions of various sorts (…) Data is the oil that powers these predictions" (Calvano and Polo, 2020). The more data they accumulate the better their knowledge of the market and the distance with competitors becomes. "Platforms can use this information asymmetry to facilitate interaction and increase welfare for users. These data externalities attract users to the platform" (Martens, 2020) triggering a circle: "The collection and use of big user data enables [platforms] to continuously improve the quality of their offerings" (Fast et al., 2021) creating network effects that "may result in monopolistic market power of platforms which they can use for their own benefit, at the expense of users" (Martens 2020).

This market evolution calls for new regulations, to better protect personal data and to ensure a level-playing field in digital markets, but this is out of the scope of this article. Here I would like to call for more collaboration within the book value chain, involving publishers, booksellers and librarians.

We, in the book community, share some key objectives. We all aim at better understanding readers' needs to offer them the best content and services. We also share some fundamental values: the respect for personal data and – above all – freedom of expression and pluralism, which, in market terms, also means fair competition and absence of monopolistic positions.

Because we share goals and values, we need to design a context where cooperation will enable all citizens and SMEs to access relevant information and intelligence derived from book-related data sets (i) at fair conditions, (ii) while respecting personal data (iii) and commercial confidentiality.

Technologies offer opportunities besides threats. The potential offered by artificial intelligence and data analysis can be exploited by the cultural sector too. In a market where network effects give immense advantage to few, cooperation among many can be the answer.

---

[8] Issue related to data exclusivity by the market "gatekeepers" are also enlightened in Recitals 43-45, 54-56 and 61. See European Commission 2020-b.

JLIS.it

References

Bell G. and Saynor G. (2018), *Thema: the Subject Category Scheme for a Global Book Trade*, Editeur: https://www.editeur.org/files/Thema/20180426%20Thema%20briefing.pdf.

Calvano E. and Polo M. (2020) Market Power, Competition and Innovation in Digital Markets: A Survey, *Information Economics and Policy*, Vol. 54, 100853, https://doi.org/10.1016/j.infoecopol.2020.100853.

European Commission (2020-a) *Making the Most of the EU's Innovative Potential. An Intellectual Property Action Plan to Support the EU's Recovery and Resilience*, COM(2020) 760 Final, 25 Nov 2020.

European Commission (2020-b), Proposal for a Regulation of the European Parliament and of the Council on Contestable and Fair Markets in the Digital Sector (Digital Markets Act), Brussels, COM (2020) 842 final, 15 Dec 2020.

European Council (2019) *Developing the Copyright Infrastructure - Stocktaking of work and progress under the Finnish Presidency*, https://data.consilium.europa.eu/doc/document/ST-15016-2019-INIT/en/pdf.

Fast V., Schnurr D. and Wohlfarth M. (2021), Regulation of Data-driven Market Power in the Digital Economy: Business Value Creation and Competitive Advantages from Big Data (January 31, 2021). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3759664.

Foer F. (2018), *World Without Mind: The Existential Threat of Big Tech*, Penguin Putnam.

IFLA International Federation of Library Associations (1997), *Functional Requirements for Bibliographic Records.* https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

Krämer J. and Wohlfarth M. (2018), Market Power, Regulatory Convergence, and the Role of Data in Digital Market, *Telecommunications Policy* Vol. 42, pp. 154–171. https://doi.org/10.1016/j.telpol.2017.10.004.

Lanier J. (2011), *You Are Not a Gadget: A Manifesto*, Penguin Books.

Lanier J. (2019), *Ten Arguments for Deleting Your Social Media Accounts Right Now*, Vintage Publishing.

Li J. (2021), A Comparative Study of Keyword Extraction Algorithms for English Texts, *Journal of Intelligent Systems*, 9 July 2021. https://doi.org/10.1515/jisys-2021-0040.

B. Martens (2020), *An Economic Perspective on Data and Platform Market Power*, JRC Digital Economy Working Paper 2020-09.

Mazzucchi P. (2021), Copyright Infrastructure: l'innovazione che fa bene alla cultura del Paese, Agenda Digitale, 9 Jun 2021.

Paskin N. (2006), Interoperability. A Report on Two Recent ISO Activities, *D-Lib Magazine*, Vol. 12, No. 4.

Rust G. and Bide M. (2000), The indecs Framework - Principles, Model and Data Dictionary. https://www.doi.org/factsheets/indecs_factsheet.html.

Vuopala A. (2021), "Copyright Infrastructure. A Recipe for Recovery and Resilience of the Creative Sectors", IPR Info, n. 2 / 2021.

JLIS.it

Two-dimensional books for the new Open Access academic publishing*

Fulvio Guatelli[a]

a) Firenze University Press, http://orcid.org/0000-0002-0309-0940

ABSTRACT

Metadata have become a key element of scientific communication. Indeed, the content of a publication – that is, what we love, discuss and judge – is no longer the alpha and omega of a scientific publication nor its exclusive centre of gravity. Books are gradually taking the form of an iceberg, whose visible part is represented by the content, while the submerged part is constituted by metadata. In the current communication approach of scientific research, metadata and dissemination go hand in hand, as metadata provide a huge contribution to the success of the research itself. In this paper, I will illustrate how – in the field of today's scholarly publishing – best practices, simple metadata, and cataloguing indicators such as DOI and ORCID are taking on the task that was once accomplished by chariots pulled by sturdy horses coming out of Aldo Manuzio's workshop: spreading books and the discoveries of scientific research all over the world.

KEYWORDS

Two-dimensional book; Open access book; Metadata; Academic publishing.

* This article is a revised and expanded version of a paper presented at the International Conference – Bibliographic Control in the Digital Ecosystem, held in Florence (Italy), on February 8th-12th 2021.

JLIS.it

I. Metadata and Scientific Communication.

Metadata is one of the most crucial topics in the educational training of cataloguers, archivists and technicians in the publishing world. For decades, metadata have accompanied books, contributing to their preservation and distribution. Until recently, however, they were just external and subsidiary elements to the scientific publication: monographs, edited volumes, and journal articles were only identified by their intelligible content, and nothing else.

Today this is no longer the case, and this article aims to describe the new scenario of scholarly publications. In this scenario, metadata have gained a new dimension, one that was unimaginable until a few years ago.

Metadata have become the protagonist of scientific communication, where a publication consists not only in its content, but also in the set of metadata associated with it. In other words, what we read, what we are passionate about or annoyed by, or bored by, what we discuss and finally evaluate, is no longer the alpha and omega of that publication, its centre of gravity. Metadata – commonly known as "the hidden data", the silent descriptive properties, or the endless tables of categories that relentlessly capture, and standardize the elusive qualities of a text – have risen to the fore.

To better convey the magnitude of this change, some facts known to the specialist as well as to the general public are worth recalling.

Let us consider, for example, the most prominent ancient Greek philosopher, Aristotle. The Philosopher is a very popular historical figure, indeed, and yet we know so little about his life. Even his contributions to human knowledge have grey areas, to the point that even his best-known book the "Physics", consisting in a collection of treatises, is a text reconstructed by his pupil Andronicus of Rhodes *a posteriori* and centuries after Aristotle's death.

However, if we had Aristotle's ORCID and the DOI of "Physics" we would have two perfectly defined entities, which could be processed by a machine capable of carrying out countless services. In other words, Aristotle is to ORCID as "Physics" is to DOI and, more or less, this is the functional strength of the so called "digital revolution". Aristotle and "Physics" possess certain intrinsic features – they are brilliant, seminal, and sometimes uncertain, obscure, as life is – while ORCID and DOI have others – they can be boring, plain, but also certain, clear, and cheap as machines are. Mutatis mutandis, it is basically the last battle of an ancient war that has involved mathematicians, physicists and philosophers and focused on continuum vs. discretum, that is the world of Continuum against the world of Discrete, truly an endless story.

As we mentioned earlier, the content of a scientific publication is no longer the sole centre of gravity of a book. Books are gradually taking the form of an iceberg, whose visible part is represented by the content, while the submerged part is constituted by metadata.

The book-iceberg association may seem an odd one, but it is not new in the field of literature. Ernest Hemingway, interviewed by George Plimpton in 1958, explained the art of fiction with these words: "I always try to write on the principle of the iceberg. There is seven-eighths of it underwater for every part that shows. Anything you know you can eliminate and it only strengthens your iceberg. It is the part that doesn't show [...]. But the knowledge is what makes the underwater part of the iceberg". (Hemingway 1958)

Moving from literature to academia, metadata and dissemination of scientific discoveries go hand in hand in the current scholarly communication approach. Metadata not only provide a huge con-

tribution to the success of research, but more importantly, they are a part of it. In Hemingway's words, they are the knowledge that makes the underwater part of the iceberg.

The transformation underway places the book and its constituent elements before several economic, social and even philosophical considerations. As a matter of facts, if the features of a given object change, the way of interacting with it also changes. Furthermore, if that object is a vehicle of human knowledge, the situation becomes exponentially more complicated and, at the same time, intriguing.
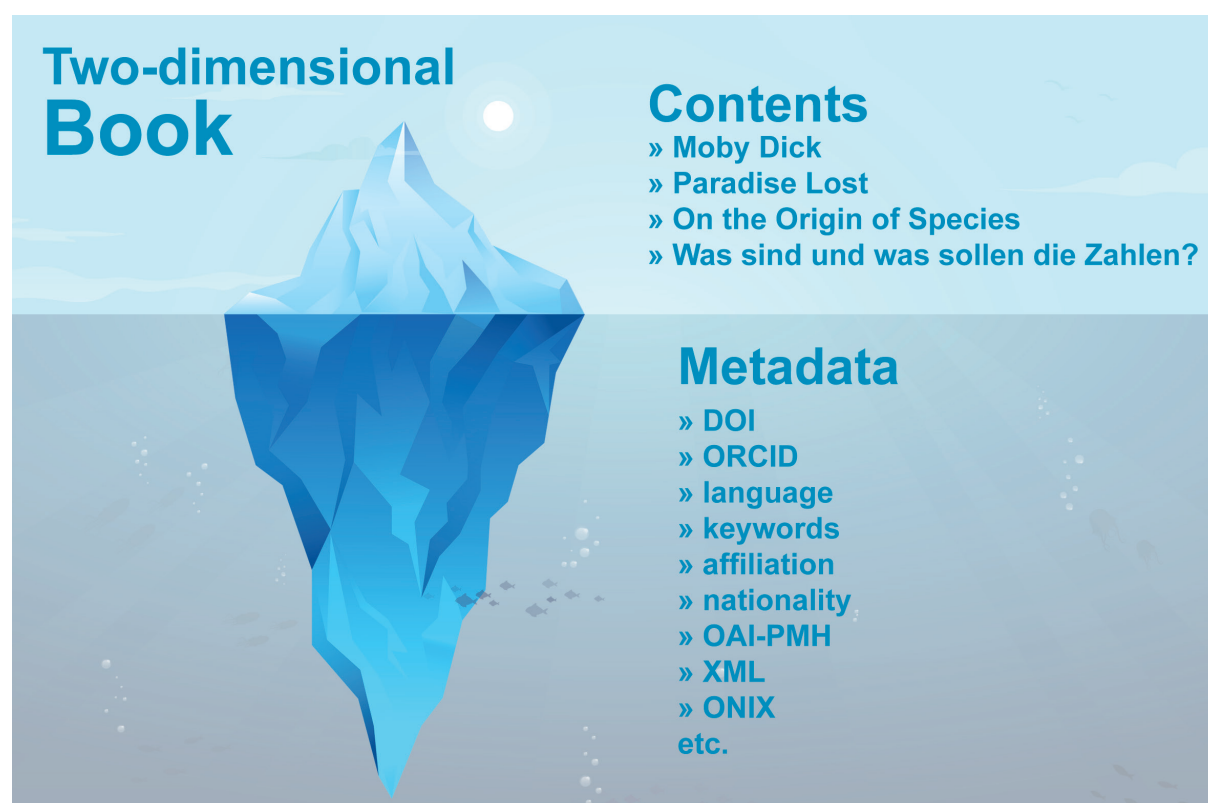
Fig. 1. Icebergs and two-dimensional books (CC BY 4.0)

II. What are Books Becoming? A Few Remarks on Research and Lifecycles

The mandatory starting point on any consideration on books nowadays is asking ourselves what the book is turning into. What seems quite clear is that, in the multifaceted academic publishing scenario, metadata, cataloguing indicators such as DOI and ORCID, and best practices – which are crucial guidelines on good scholarly publications – will increasingly play a significant role in the creation of a book (Adema and Stone, 2017; Capaccioni 2014).

Getting more specific, the new shape of books – in which content and metadata are bonded together like the two sides of the same iceberg – is becoming more and more embedded in the research lifecycle. As scholars experience every day, the research lifecycle consists of various stages, the main being: Planning and Funding, Conducting Research, Considering Publishing Options, Writing and Submitting the Manuscript, Peer Review, Publishing Contract and License, Publish-

JLIS.it

ing and Dissemination, Reuse of Research. All together, these stages represent the lifecycle of any research.

Regardless of the disciplinary fields (HSS or STM), of cultural traditions, of the scholar being a scientist in a large research group working under a mountain chasing down subatomic particles, or a philologist working alone among manuscripts looking for Machiavelli unpublished works, the result will be the same. To be active agents of the new scenario of scientific communication, books must be fully embedded in the research lifecycle featured above. Basically, this transformation is already happening, right now.
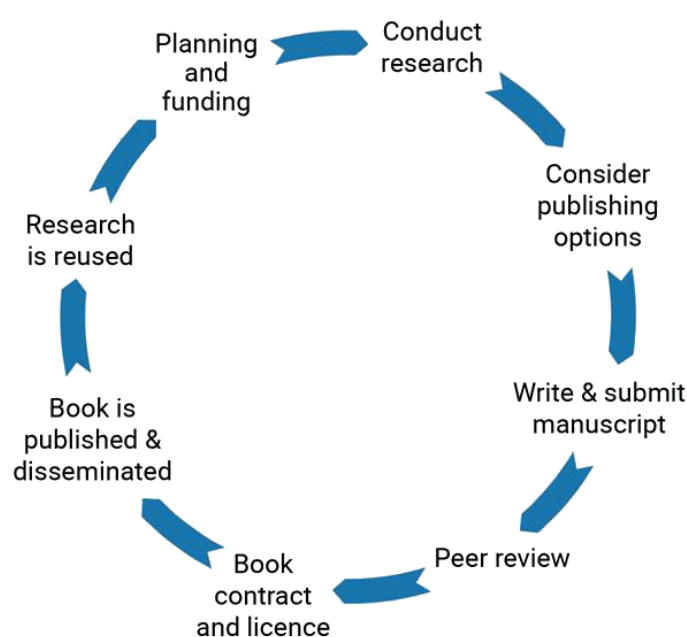
Fig. 2. Research lifecycle (OAPEN OA Books Toolkit, CC BY 4.0)

Observing the process from a practical point of view, what does it mean that books are getting more embedded in the research lifecycle? To answer this question, a closer look at another cycle will help, that of the publishing lifecycle. At the origin of a scholarly work such as a monograph, a research is proposed, funded, and reported on. Then the monograph is evaluated to assess its quality, and it is edited by peers. After that, a publisher provides editing, layout, and publication services, and the work is published. Then, it is disseminated according to a well-defined access model. Works are distributed in print or online, through libraries, retailers, and on the Web, and preserved (copies or versions of the work may be saved for posterity). Obviously, the work is then reused in a constant evolving lifecycle (works get read, cited, and recombined).
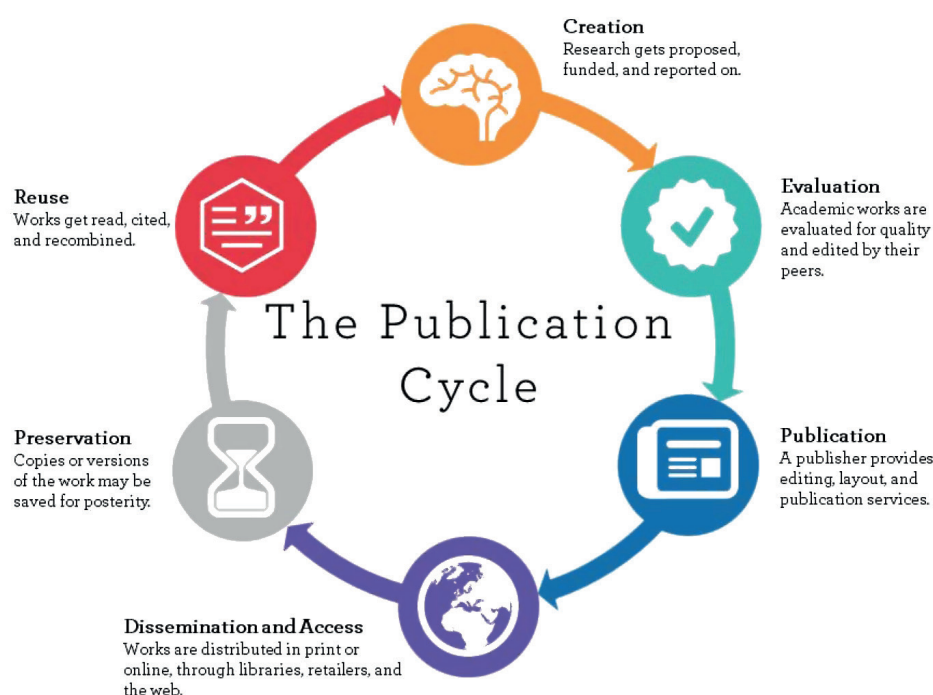
JLIS.it

Fig. 3. Publication lifecycle (Berkeley Library Scholarly Communication Services, CC BY-NC 4.0)

The iceberg-volume may interact in an unprecedent way with the entire publication lifecycle, where creation, evaluation, dissemination, access, and preservation are not actions performed around the book, but rather key features of the book itself. By linking the research lifecycle and the publishing lifecycle, digital books have the potential to innovate the whole scenario, by providing new tools and solutions in the four main areas of a publication, namely: (i) authorship, (ii) publishing formats, (iii) evaluation process, and (iv)access, considered as dissemination and impact.

The "FUP Scientific Cloud for Books" project provides with a suitable example on such topic, since it was conceived to develop a model of working practices that would ensure the production of the new generation monographs described in this article.

Launched in 2019 by the Firenze University Press, the project has captured the ongoing change of books with the aim to increase the dissemination and impact rates of its monographic publications (Guerrini and Ventura 2009). In a communicative landscape in which metadata and the dissemination of scientific discoveries go hand in hand, metadata become co-responsible for the success of a scientific publication. The project was also aiming at filling the gap existing between scientific journals, where digital has enhanced both visibility and impact, and the monograph (British Academy 2018, 2019; Guatelli and Pierno 2015, 85–113). It is a matter of fact that the latter, while representing a fundamental tool for academic dissemination and career progression, is still a rather marginal player in the digital revolution.

The project aims at providing a systematic and thorough attribution of machine-readable metadata and formats (Guatelli 2018, 2020, 47–57). Such attribution applies to all the four key areas of a publication, as already mentioned. Therefore, any digital volume must meet the following standards:

- Authorship: all the authorial components of a volume must be identified by a defined set of metadata. Therefore, the authors of books or single chapters, editors, but also the those involved in the evaluation process (such as editor-in-chiefs, members of scientific boards, referees, research institutions and funders) are systematically described by using simple but effective metadata: first/last name, affiliation, nationality, ORCID, e-mail;

- Publication formats: volumes are currently published in multi-format editions. These can be, for instance, PDF, epub, html, or xml. Particular emphasis must be put on machine-readable formats, as they are functional both to machine-learning processes and information retrieval (IR) systems, and to the processes of dissemination through indexes and aggregators, such as DOAB, OAPEN, WorldCat, OAlster, ProQuest, EBSCO, SBART, OPENAIR, etc.

- Evaluation: each volume must clearly report the characteristics of the evaluation process to which it has been subject. The scope here is a wide one, as it includes references to the applied best practices (namely, peer review policy, open access policy, copyright and licensing policy, publication ethics and complaint policy, e.g. https://fupress.com/fup-best-practice-in-scholarly-publishing) and to the referee list of the book series, also providing the reader with basic statistical data on the refereeing process (date of paper submission, date of acceptance, and the like).

- Access, dissemination and impact: among the four areas, the innovations related to access, dissemination and impact are particularly remarkable and deserve further analysis:

    i) Open Access: Firenze University Press fully supports Open Access publishing as it is an exceptional tool to share ideas and knowledge in all disciplines with an open, collaborative, and non-profit approach (Delle Donne 2010, 125–50; 2018; European Commission 2019; Ferwerda, Pinter, Stern, and Niels 2017). Open Access books and book chapters allow the research community to achieve wide and rapid dissemination across all book formats, as well as a high impact for their research. All FUP content and metadata are published in Open Access, released under Creative Commons licenses stating the Author as the copyright holder (https://fupress.com/open-access-copyright-and-licensing-policy).

    ii) Dissemination: to increase discoverability, access and shareability of peer-reviewed research, the publisher endeavours an ongoing activity of indexing of its books and book chapters on dedicated platforms for hosting, dissemination, discovery, and preservation. It supports and encourages research libraries, as well as profit and non-profit indexing services, to list its series, books and book chapters among their electronic resources. All our book metadata are openly available for download in various formats by any indexing service (OAI-PMH, XML, etc). Metadata are released under the Public Domain Dedication license (CC0 1.0) (eg. https://fupress.com/distributions-indexing-and-abstracting-policy).

    iii) Impact: For each book and book chapter published, Firenze University Press provides the author with periodically updated usage statistics (about books and book chapters downloads and views) according to the international standard currently used in positioning and evaluation processes (the COUNTER Code of Practice for Release 5 standard).

JLIS.it

By applying the formula briefly summarized above, the resulting editorial product becomes an innovative digital book featuring a deep interaction between content and metadata. Monographs implemented in this way can ensure high indexes of dissemination, filling the gap with scientific journals that used to have an edge in the area of impact until recently (Vincent 2013, 107–119; Gatti and Mierowsky 2016, 456–59; Neylon, Montgomery, Ozaygen, Saunders, and Pinter 2018). To use a charming and historical example, best practices, metadata and cataloguing indicators, such as DOI and ORCID (Jisc 2018; Tsuji 2018; UK Research and Innovation 2020), are taking on the task that was once accomplished by chariots pulled by sturdy horses coming out of Aldo Manuzio's workshop: spreading books and the discoveries of scientific research all over the world. The iceberg-book approach promoted and realized within the framework of the "FUP Scientific Cloud for Books", however, is not limited to enhancing dissemination; rather, its innovative approach consists in expanding the identity of the book in its two dimensions, under and over the ocean. This is the real strength of such an approach.

Born as a pioneering experiment, the project is yielding greater fruits than the most optimistic forecasts, even hinting at potential further development. The revolution behind the iceberg-book is somehow reminiscent of both the cathedral and the bazaar described by Eric Raymond in his famous essay (Raymond 1999). In software development, the author described two models, one closed and verticalized, the cathedral, and one open to user interaction, the bazaar. The new digital book preserves both verticalization and closure (the book always has an author and specific "boundaries") and the participation of different subjects, both in production and in open access fruition. Its open and shareable part is only at the beginning of a transformation process that could one day turn readers as well into active subjects in the certification/dissemination of monographs. As has recently been pointed out on open access (Capaccioni 2019), one must keep in mind that scholarly communication is always a space within which different actors act and are all relevant. Speaking of the future inclusion of readers in the process, we do not know what will eventually happen to the iceberg, but it will be extremely inspiring to watch it unfold.
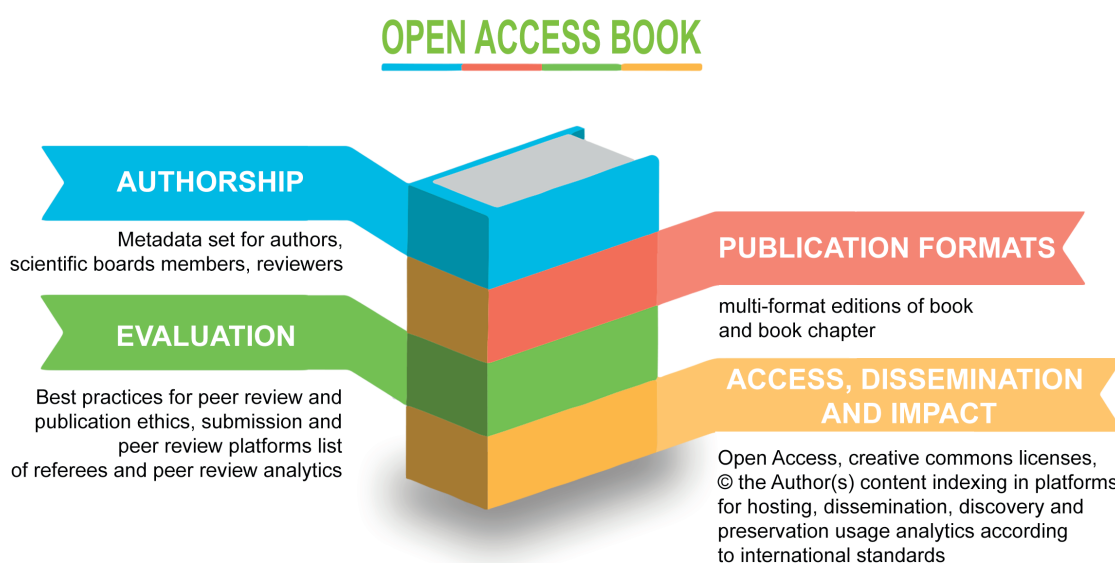

OPEN ACCESS BOOK

AUTHORSHIP
Metadata set for authors, scientific boards members, reviewers

PUBLICATION FORMATS
multi-format editions of book and book chapter

EVALUATION
Best practices for peer review and publication ethics, submission and peer review platforms list of referees and peer review analytics

ACCESS, DISSEMINATION AND IMPACT
Open Access, creative commons licenses, © the Author(s) content indexing in platforms for hosting, dissemination, discovery and preservation usage analytics according to international standards

Fig. 4. Open Access Books (CC BY 4.0)

JLIS.it

References

Adema, Janneke, and Graham Stone. 2017. *Changing publishing ecologies. A landscape study of new university presses and academic-led publishing, a report to Jisc.* http://repository.jisc.ac.uk/6666/1/Changing-publishing-ecologies-report.pdf.

British Academy. 2019. *Open Access and Book Chapters. A report from the British Academy* https://www.thebritishacademy.ac.uk/publications/open-access-book-chapters-report/

British Academy. 2018. *Open access and monographs: Where are we now?* https://www.thebritishacademy.ac.uk/publications/open-access-monographs-where-are-we-now/

Capaccioni, Andrea. 2014. "La monografia scientifica e le sfide dell'accesso aperto." *AIB Studi* 54, 2/3: 201–11. https://doi.org/10.2426/aibstudi-10084.

Capaccioni, Andrea. 2019. "La monografia ad accesso aperto e gli sviluppi dell'Open Access". *JLIS.it* 10: 59–71. http://dx.doi.org/10.4403/jlis.it-12516.

Delle Donne, Roberto. 2018. "L'accesso aperto, le università e le SSH." *Il Capitale culturale. Studies on the Value of Cultural Heritage* 17: 17–45. http://doi.org/10.13138/2039-2362/1944.

Delle Donne, Roberto. 2010. "Open access e pratiche della comunicazione scientifica. Le politiche della CRUI." In *Gli archivi istituzionali. Open access, valutazione della ricerca e diritto d'autore*, edited by Mauro Guerrini, 125–50. Milan: Editrice bibliografica.

European Commission. 2019. *Trends for open access to publications* https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/trends-open-access-publications_en

Ferwerda, Eelco, Frances Pinter, and Niels Stern. 2017. *A landscape study on Open Access and monographs: policies, funding and publishing in eight European countries.* http://doi.org/10.5281/zenodo.815932.

Firenze University Press Best Practice in Scholarly Publishing. https://doi.org/10.36253/fup_best_practice

Gatti, Rupert, and Marc Mierowsky. 2016. "Funding open access monographs. A coalition of libraries and publishers." *College & Research Libraries News* 77, 9: 456–59. https://crln.acrl.org/index.php/crlnews/article/view/9557/10902.

Guatelli, Fulvio. 2020. "FUP Scientific Cloud e l'editoria fatta dagli studiosi", *Società e Storia* 167: 155–64. http://dx.doi.org/10.3280/SS2020-167008.

Guatelli, Fulvio. 2018, *Editoria, università e la nuova "comedia": riflessioni sul ruolo delle istituzioni di ricerca nella disseminazione della scienza*, Il Capitale culturale. Studies on the Value of Cultural Heritage, 17: 47–57, http://dx.doi.org/10.13138/2039-2362/1901.

Guatelli, Fulvio, and Alessandro Pierno. 2015. "*Pubblicare open access journal: dalla progettazione alla promozione.*" In *Via verde e via d'oro. Le politiche open access dell'Università di Firenze*, edited by Mauro Guerrini, and Giovanni Mari, 85–113. Florence: Firenze University Press. <https://fupress.com/catalogo/via-verde-e-via-d'oro/2873>, 24.03.2018.

JLIS.it

Guerrini, Mauro, and Roberto Ventura. 2009. "Problemi dell'editoria universitaria oggi: il ruolo delle university press e il movimento a favore dell'open access". In *Dalla pecia all'e-book: libri per l'università: stampa, editoria, circolazione e lettura. Atti del convegno internazionale di studi (Bologna, 21-25 ottobre 2008)*, edited by Gian Paolo Brizzi, and Maria Gioia Tavoni, 665–70, Bologna: CLUEB.

Hemingway, Ernest. 1958. "The Art of Fiction." Interviewed by George Plimpton. *Paris Review*, No. 21, Issue 18, Spring.

Jisc. 2018. *Open Access Monographs in the UK* https://repository.jisc.ac.uk/7090/1/2018JiscOABriefingOAMonographsUK.pdf

Neylon, Cameron, Montgomery, Lucy, Ozaygen, Alkim, Saunders, Neil, Pinter, Frances. 2018. *The visibility of Open Access monographs in a European context: full report.* http://doi.org/10.5281/zenodo.1230342.

Raymond, Eric Steven. 1999. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary.* Sebastopol, CA, USA: O'Reilly Media.

Tsuji, Keita. 2018. "Statistics on Open Access Books Available through the Directory of Open Access Books." *International Journal of Academic Library and Information Science* 6, 4: 86–100, http://doi.org/10.14662/IJALIS2018.031, available from: https://arxiv.org/pdf/1808.01541.pdf.

UK Research and Innovation. 2020. *UKRI Open Access Review: Consultation*, https://www.ukri.org/wp-content/uploads/2020/10/UKRI-231020-OpenAccessReview-Consultation25Mar20.pdf

Vincent, Nigel. 2013. "The monograph challenge." In *Debating Open Access*, edited by Nigel Vincent and Chris Wickham, 107–119. London: British Academy. https://www.britac.ac.uk/sites/default/files/Debating-Open-Access-2013.pdf.

JLIS.it

Bibliographic control and institutional repositories: welcome to the jungle

Tessa Piazzini[(a)]

a) Università degli Studi di Firenze, http://orcid.org/0000-0002-8876-371X

ABSTRACT

In 1994 cognitive scientist Stevan Harnad made what he defined a "subversive proposal" to his colleagues: «immediately start serf-archiving their papers on the Internet». Since then, institutional repositories have been chaotically developing, alongside disciplinary repositories. In the early XXI Century the public debate was centered on their purposes and therefore on what they were supposed to contain; librarians joined the discussion and contributed to it by implementing descriptive standards such as Dublin Core and interoperability protocols (OAI-PMH). The themes under discussion were closely related to bibliographic and authority control, given that the quality of metadata has a profound impact on the quality of the services offered to users. Presently, we are still trying to answer some of those old questions: what (or whom) are IRs for? Is bibliographic control so necessary within an environment that has never failed in self-archiving? Can we consider IRs a bibliographic tool? We also need to deal with a wider vision: in a scenario that saw the transition from OPACs (created, managed and controlled by librarians) to current discovery tools (with their information redundancy and the related problems on data correctness and quality control) can librarians still be authoritative and act effectively?

KEYWORDS

Authority control; Institutional repository; Bibliographic control; Metadata.

JLIS.it

1. Introduction

In library sciences we often talk about "ecosystems"[1]. Within this naturalistic metaphor, I used to think of repositories as a jungle: chaotic, dense and impenetrable. Now, however, I look at repositories rather as a rain forest: an equally complex environment full of variety, multi-layered, characterized by a lot of internal and external communication networks and also hidden and visible interdependencies. An environment that rests on a clean, rich soil, on which it is possible to move and walk.

The relatively short history of repositories shows us a great variety in terms of stored material (pre-print, research publications, teaching materials, articles and books, theses, multidisciplinary or specialized), in terms of population and organization (self-archiving, batch retrieval, internal collections, librarian mediated insertion) and also in terms of software (Digital Commons, DSpace, Eprints, Fedora…) .

We must not forget that the repositories were born because of the initiative of the scientific community, in particular by the will of single authors; it all started with the "subversive" proposal by Stevan Harnad, professor of cognitive sciences at the Virginia Polytechnic Institute, in 1994: sharing their own research within the institution through the self-archiving of online contributions, in order to make their dissemination more effective. Hence the embryo of a new type of open archive: the institutional archive, promoted and managed by an institution, which goes alongside the disciplinary ones created by aggregation of documents concerning single research areas. We should also remember that "the environment and context in which a repository is situated will unequivocally play a part in the decisions that are made and the quality of metadata that is produced" (Moulaison Sandy and Dykas 2016, 105).

Such repositories, therefore, are fed by the authors themselves through the practice of "self-archiving". It is no coincidence that the elements on which the foundations of repositories rest are the use of minimal metadata sets (such as the Dublin Core Metadata Element Set, adopted by most repository providers) and interoperability standards, such as the OAI-PMH , which go hand in hand with such minimality. According to some scholars, this choice was due to the desire of encouraging the participation of authors in filling the repositories, however it did discourage any activity that could be perceived as a barrier between scholars and their own institutions; this perception has also affected any intervention and quality control measure in the process of metadata creation (Barton, Currier and Hey 2003). Consequently, unlike other more familiar environments, such as OPACs and Discovery Tools, institutional repositories were not created by librarians to organize and make available the bibliographic universe; they are not tools whose birth falls within our sphere, although the contribution of many librarians has undoubtedly been, and still is, vast (for example, in the case of the development of the Dublin Core, despite the "bibliographic control community" seeing it at its birth as "simplistic"[2]).

The awareness that "[for IRs] there is no universally accepted practice or standard defining quality

---

[1] In this issue many speakers have included this word in the titles of their articles.

[2] In his keynote address to the Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium, Michael Gorman (2001, xxv) referred to metadata as "a fancy name for an inferior form of cataloguing," and as "unstandardized, uncontrolled, ersatz cataloging." Cited in footnote 6 by Howarth 2005, Metadata and Bibliographic control: soul mates o two solitudes, Cataloging and classification Quarterly, 40 (3-4), 37-56.

JLIS.it

metadata; similarly, there is no set of rules for describing institutional repository materials" (Stein, Applegate and Robbins 2017, 650) marks even more the difference between IRs and actual bibliographic tools.

2. Repositories as rain forests or woodlands?

This awareness leads to a first question: what should we do about these environments? Should we maintain their relatively unorganized nature (with the relative entropy) or should we transform these rain forests into controlled, orderly and organized woodlands, according to our vision of the (bibliographic) world?

Maybe we should be looking for the golden middle way, or the "aurea mediocritas", to quote the Latin poet Horace, who reminded us "est modus in rebus": we should tread carefully and with the respect that is due to an environment with its own characteristics; we should try to intervene most discretely, in order not to deform it, but simply to provide an orientation to its users, so that they not get lost and can appreciate all its beauty.

Therefore, perhaps bibliographic and authority control should also be careful, aiming at providing the users (both humans and machines) with the necessary information in the best possible form, for them to enjoy a pleasant, safe and satisfying journey.

Although born within the academic communities and made their own by the institutions, the repositories, in fact, cater to the widest possible audience, and their main goal is – desired by the authors themselves – not so much to organize and systematize what is produced by an institution (a task that can well be carried out by a catalog), but to ensure the widest visibility for the longest time.

Long-term access and storage can be achieved through two closely linked elements: good metadata and a good repository system; in fact "quality metadata may be underutilized due to weakness in indexing, navigation, and display options"(Moulaison Sandy and Dykas 2016, 103).

Defining what is meant by "quality metadata" is not, however, an easy task in itself: the subjective and local elements remain strong, and the close link between functional requirements and suitability for purpose is often emphasized (Powell, Day and Guy 2004) "by defining both the internal requirements related to the needs of end-users in a local setting and by defining external requirements related to disclose and expose local metadata relating to external service providers" (Park 2009, 214).

At the same time, therefore, it is necessary to guarantee interoperability within a Search Engine Optimization framework, and that implies a much wider series of technical activities and operational choices that should be part of the development plan for a repository.

As librarians we have the task of providing support for the creation of quality metadata, as the NISO (National Information Standards Organization) reminds us in its *A Framework of Guidance for Building Good Digital Collections* (NISO 2007, 61-62), when it says that Good Metadata:

1. Conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and potential future uses of the collection.
2. Supports interoperability.
3. Uses authority control and content standards to describe objects and collocate related objects.
4. Includes a clear statement of the conditions and terms of use for the digital objects.

5. Supports the long-term curation and preservation of objects in collection.
6. Good metadata records are objects themselves and therefore should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification.

When speaking of quality of metadata, we cannot, therefore, neglect a fourth aspect besides those of in accuracy, completeness and consistency, which we always find in the literature (Park 2009): their effectiveness in terms of information retrieval, in particular in relation to indexing by search engines. A 2019 study (Mering 2019) conducted upon the University of Nebraska-Lincoln IR tells us that 57% of the collection was accessed via Google Search and only 17% was accessed directly from within the repository. Already in 2012 Arlitsch and O'Brien reminded us that "digital repositories [...] face a common challenge: having their content found by interested users in a crowded sea of information on the Internet" (Arlitsch and O'Brien 2012, 64). Their research also showed that all the analyzed repositories, regardless of the platform, had a low index ratio making them virtually invisible to Google Scholar.

The ultimate goal, however, remains that of ensuring the widest visibility (Swan and Carr, 2008), which is also achieved by making the repository content externally shareable. "Shareability implies an adherence to internal but also extra-organizational standards and best practices; when every repository uses recommended file types, metadata schemas, and the same controlled vocabularies, information is more easily searched and retrieved across them" (Moulaison Sandy and Dykas 2016, 102).

It is therefore important that "for metadata to be effective, enforcement of standards of quality must take place at the community level" (Bruce and Hillman 2004, 3) and that it is necessary to "establish effective policies for the management of authorities in these types of digital collection through cooperative efforts that will permit the development of corpora of authority entries that will aid the processes of cataloging, metadata creation and information retrieval" (Barrionuevo Almuzara, Alvite Díez and Rodríguez Bravo 2012, 101).

Economics, however, teaches us that effectiveness, in cost-benefit analysis, is always accompanied by efficiency, while our daily experience as librarians continually reminds us that bibliographic control activities are very time-consuming and resource-consuming and also require high professional skills.

For example, a survey (Moulaison Sandy and Dykas 2016), conducted in 2015 among some US repository administrators, indicated time limitations and staff hours and skills levels of staff among the greatest obstacles to the provision of high quality metadata, even for those who had self-assessed the quality of their own metadata repository as above average.

3. Different levels of supports for authors and staff in bibliographic control

Hence the need to make prudent choices on how to distribute efforts and how to use all the tools that can help reduce costs.

Now, in the face of the continuous growth of scientific publications – a result also of increasingly pervasive evaluation activities – especially when the main source is self-archiving, one of the tasks of librarians within IRs is to build clear and fast paths, which simultaneously guarantee ease of

use and quality of information both for authors and readers, by working on those metadata that have an impact on access.

The first basic form of support consists in offering authors or editors explanatory notes, instructions, drop-down menus or pop-up windows when filling in the bibliographic form online: in this case we certainly cannot speak of bibliographic control; nevertheless, it is a first step which we cannot ignore. Providing clear information, simple tools and technical support goes along with training and education activities, and they all constitute an indispensable step towards creating a community of users who have awareness of the creation of quality metadata, as to reduce *ex ante* the need for subsequent controls.

At the same, initial, layer of support we find the preliminary definition of best practices and guidelines destined to repository managers and staff: "metadata guidelines seem to be fundamental in ensuring a minimum level of consistency in resource description within a collection and across distributed digital repositories" (Park and Tosaka 2010, 711).

A second level consists of recovering quality data from third-party sources through identifiers.

Here is where another naturalistic (or rather, environmentalist) metaphor comes into play: recycling, intended as the possibility of quickly recovering structured and valid information from external sources.

Currently, when it comes to bibliographic and authority control, scholars seem to agree on two main tools:

- Unique publication identifiers for retrieving verified metadata;
- Unique personal identifiers (ORCID, VIAF, ISNI, LoC Authority Name) to ensure the quality control of the authors.

Both tools currently have some limitations and difficulties, but there seems to be a certain agreement among scholars on their effectiveness.

With an eye to economic sustainability, there is a growing tendency for repositories to offer the possibility to retrieve a lot of information from external databases or directly from the publisher through DOI, Scopus identifier or Web of Science Accession number, ISBN, PMID.

A choice of this kind implies on our part, as librarians, the acceptance of delegating the organization and presentation of data to third parties, but this is not new to us: we have done it with derived cataloging, we do it today in part with discovery tools.

If, on the one hand, this can in principle lead to a certain homogeneity in the presentation of information, reducing the risk of typos and errors deriving from manual entry, on the other hand it only partially reduces the need for bibliographic control. This is first of all because publishers and individual databases each have their own metadata and cataloging rules, and secondly because harvesting activities strictly depend on the interoperability protocols applied and on the mapping between the different sets of metadata (Chapman, Reynolds and Shreeves 2009).

For example: the title of an article in the Pubmed database is always presented in English, even if the article is published in another language. In this form the dc.title field of the Core Set Dublin Core is usually imported, where, instead, there should be the title in the original language.

Furthermore, we know well how partial in their coverage large international databases are when it comes to languages or subject matters, and how there are no specific identifiers for certain kinds of publications such as, for example, the essay within a volume – although many publishers are beginning to equip individual book chapters with their own DOI.

Even in the case of batch uploads carried out by dedicated staff, the quality of the data is not guaranteed, indeed in some cases it seems to be even lower (Stein, Applegate and Robbins 2017), unless an intense metadata cleanup is planned prior to batch ingestion using tools like Open Refine[3].

4. Authority control as part of bibliographic control

We find a similar problem especially in the management of authors' names, whose ever-changing forms have always been a great challenge for the authority control. It could be due to the will of authors (discontinuous use of the middle name or of abbreviated forms, change of surname after marriage...) or for other reasons (different presentations in various sources, linguistic variants, transliterations, etc.).

In order to overcome this problem, as already noted, a first form of assistance is the auto-completion function that – although not present in every relevant software – can kick in during the insertion phase: although this functionality can contribute to the reduction of variant forms, this does not strengthen the authority control.

At a slightly higher level – so much so that we can speak of authority control at a local level – we find the automatic linking of the author's name to the institution identity management system; nevertheless it is evident that this solution does not fully guarantee the interoperability and shareability and "can be, at best, [only] one part of the repository authority control puzzle" (Downey 2019, 130).

In order to achieve this result, unique identifiers of the names are being implemented within the repositories, by linking to external authority schemes.

The interesting aspect, determined by the evolution towards an Linked Open Data model, is the transition from the concept of "name authority work" to that of "identity management", thanks to the association of registered identifiers: "Identity management won't work the same way as the traditional authority control because identity management emphasizes the process of associating a registered identifier (or a URI) with a single entity and the differentiation of names or headings is only of secondary importance in identity management " (Zhu 2019, 227). The coexistence, however, of numerous projects for the name authority control with the consequent production of different identifiers constitutes an additional element of noise and can lead to the necessity, once again, to make choices.

In 2019 Moira Downey, a colleague from Duke University, published an analysis (Downey 2019) of three among the major international authority sources – Library of Congress Name Authority Files (LCNAF), Virtual International Authority File (VIAF), and Open Researcher and Contributor Identifier (ORCID) – looking to develop a Linked Data authority control within their institutional repository, given the ability of the mentioned systems to provide author URI's via API.

According to the author, LCNAF and above all VIAF, which has developed a cooperative model for the aggregation of authority data from national and regional sources with an intense activity of clustering, merging and deduplication, constitute "a broader step forward in preparing library data

---

[3] Problem also reported for the management of name entries by Salo Dorothea, 2009. "Name authority control in Institutional repositories", *Cataloging and classification quarterly*, 47 (3-4), 249-261. doi: 10.1080/01639370902737232

JLIS.it

for better integration with the broader web" (Ibid., 120), but still rely on traditional mechanisms of participatory cataloging and authority control that have an impact on the creation of identifiers, in particular for authors of articles in journals, who happen to represent the category with the biggest presence within many academic repositories, given the hyperproduction of literature in the fields of medical sciences and STEM (Science, Technology, Engineering and Mathematics). Nevertheless, they guarantee reliability on their persistence, thanks to the professionals involved, with the creation of "record in structured, machine-actionable format that did not require additional resources or inferences to ascertain" (Ibid., 131).

ORCID, on the other hand, operates as a "self claim researcher registry", which seems to delegate the authority control traditionally carried out by libraries directly to researchers, also giving them a certain autonomy of choice on their online identity in the universe of academic communication. A leap of faith by librarians or the recognition that we can no longer be the absolute rulers of the organization and presentation of information?

As often happens, the reality lies somewhere in between: ORCID URI's prove to be a good solution for self-archiving, but "the undifferentiated nature of the current ORCID database system seems unhelpful for bulk remediation of existing repository content or for large scale batch operations" (Ibid., 131).

In particular, we have no certainty about the persistence of this identifier, given that any author could decide to remove their profile at any time. We also have the same problem with local registries, which by nature are closely linked to the duration of the author's presence within the institution.

I believe that offering authors a tool, even if not a perfect one, to present themselves is a courageous and intellectually honest choice that we can support by making its use and implementation as easy as possible within "our" repositories, and continuing to invest in a parallel education and information activity on best practices.

There are now numerous experiences in this sense in the world.

For example, in Italy, in 2015, during the national research quality assessment exercise (VQR 2011-2014), the IRIDE project (Italian Research IDentifier for Evaluation) was launched, aiming to equip the Italian academic community (professors, university researchers and research institutions, doctoral students and post-docs) with a persistent ORCID identifier, by activating the registration procedure directly within the repository of their institution. Most of the Italian university repositories, which use proprietary software, have since then allowed, through a push and pull system, a bidirectional communication with their ORCID account.

Equally interesting is the experience (Svantesson and Steletti 2019) of the European University Institute of Fiesole (one of the hosts of 2021 Florence Conference on Bibliographic Control in the Digital Ecosystem) for the integration of its databases – CADMUS, EUI Central Persons Registry – with ORCID, which is also a solution for the authority control over the names of authors that partially compensates for the absence of a CRIS (Current Research Information System). The choice of using the form in the repository as the preferred name is particularly interesting, reminding us that "the criteria of selecting which of the various IDs to use will depend on the stakeholder. Among the factors to be considered is to select the ID system which attracts the "critical mass" representing one's peers" (Smith-Yoshimura et al. 2014, 9).

JLIS.it

5. The challenge of the semantic control

If there is a certain agreement on pursuing these paths, that is not the case regarding the opportunity to invest time and resources for bibliographic control on the semantic component.

In the context of an institutional repository, in many cases a multidisciplinary one, often fed by the authors themselves, the depth, breadth and variety of disciplines means that the use of subject-controlled terms is possible only at a high level, if we are to maintain homogeneity within the repository itself.

If, on the other hand, we want to respect the heterogeneity and we let the communities self-discipline, we will end up with a repository in which the consistency of semantic metadata will be extremely varied: from their total absence to populating via recognized thesauri, and in between the complexity determined by the use of synonyms, homonyms and grammar, spelling and linguistic variants.

We are again faced with the entropy vs. control dilemma: how far must our intervention as repository managers go? Lubas, speaking of PhD theses and dissertation repositories, argued that any subject indexing intervention by staff should have complemented and not replaced the choice of keywords made by the authors, even if this would have led to an inevitable increase in noise (Lubas 2009). A normalization of the keywords, or their mapping on a pre-existing controlled vocabulary would have, in fact, eliminated the unique perspective with which authors refer to their work (Radio 2014).

An interesting study from 2018 (White, Chen and Liu, 2018) tried to analyze the relationship between the presence of some metadata in the Duke University Law School repository and the number of downloads, to understand the effectiveness of the metadata itself.

The results were surprising: the number of co-authors and the presence and the number of keywords (whether they were free text or derived from controlled vocabularies) had a substantially negative correlation with downloads and were not essential for users to reach the content. On the contrary, the presence and length of the abstract had a significantly positive impact.

This partly contradicts our certainty about the importance of subject indexing and its effectiveness. A certainty already undermined over the years by studies which invited us to abandon their use (Calhoun 2006), even within the catalogs. At the same time, other studies confirmed its efficacy (Gross, Taylor and Joudrey 2015). This polarization has now widened with the adoption of Discovery tools by many libraries, which reproduce a sort of "Google-like" environment – in the name of an alleged desire to make the information retrieval experience as satisfying as possible for the user – without, however, the power of Google's Page Rank. Perhaps the term "jungle" is better suited to define such tools, more than IRs.

In order to ensure a balance, a great help could come from Linked Open Data and the Semantic Web, which, with regard to subject indexing, could make an important contribution to the enrichment of contents and bibliographic control, through a simpler management of multiple languages, better linkability of resources and simpler reuse of authority registries in applications. Furthermore, the "semantic search enables a new set of queries that are based on the power of inference engines and are not possible with traditional keyword based search" (Solomou and Koutsomitropoulos 2015, 66).

While repositories go through an inevitable initial effort of adaptation, the choice between differ-

JLIS.it

ent technologies depends on the complexity and rigor required by the specific environment (Zhu 2019). As already seen at the beginning for quality metadata, also when choosing semantic web tools the relationship between suitability for the purpose and the peculiarities of the environment to which the tools are applied must be addressed, allowing a possible scalarity of choice.

There are also many interesting experiences in this area: in 2017, at the Central University of Gujarat, India, a prototype (Khumar 2018) was developed, in which they linked the Dbpedia knowledge base to a Dspace-based repository, chosen as a linked dataset for its broad disciplinary coverage, for its automated updating mechanisms and its multilingual information support. Equally interesting is the research conducted by Greek scholars on "a transformation engine which can be used to convert an existing institutional repository installation into a Linked Open Data repository: the data that exist in a DSpace repository can be semantically annotated to serve as a Semantic Web (meta) data repository " (Konstantinou, Spanos, Houssos and Mitrou 2013, 834). And it's not the only research of this kind[4].

6. Conclusion

At the end of this absolutely non-exhaustive overview, the conclusion is that there are more questions than answers, more doubts than certainties.

However, it is clear that librarians will not be able to fail in their task of helping to build reliable, rich and "clean" repositories, while exploiting the potential offered by third parties for the creation of quality metadata and bibliographic control.

There are many roads that are being tried to build quality repositories, in which bibliographic control is effective and functional: some will be dead ends, others will become well-marked paths through which librarians and users will be able enjoy the rich rainforest that institutional repositories represent.

All we need to do is to keep on exploring.

Acknowledgements

---

[4] See also H. Fari, S. Khan and MY Javed, "Publishing institutional repositories metadata on the semantic web," *Eighth International Conference on Digital Information Management (ICDIM 2013)*, Islamabad, 2013, 79-84, DOI: https://dx.doi.org/10.1109/ICDIM.2013.6694008 and Robert J. Hilliker, Melanie Wacker and Amy L. Nurnberger 2013. "Improving Discovery of and Access to Digital Repository Contents Using Semantic Web Standards: Columbia University's Academic Commons", *Journal of Library Metadata*, 13(2-3), 80-94, DOI: https://doi.org/10.1080/19386389.2013.826036

JLIS.it

References

Almuzara Barrionuevo, Leticia, Maria Luisa Díez Alvite, and Blanca Rodríguez Bravo. 2012. "A Study Of Authority Control in Spanish University Repositories." *Knowledge Organization* 39 (2): 95-103. https://doi.org/10.5771/0943-7444-2012-2-95-1

Arlitsch, Kenning, and Patrick S. O'Brien. 2012. "Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar." *Library Hi Tech* 30 (1): 60-81. https://doi.org/10.1108/07378831211213210

Barton, Jane, Sarah Currier, and Jessie M. N. Hey. 2003. "Building Quality Assurance into Metadata Creation: An Analysis based on the Learning Objects and e-Prints Communities of Practice." *International Conference on Dublin Core and Metadata Applications; DC-2003--Seattle Proceedings*. Accessed 22 November 2021. https://dcpapers.dublincore.org/pubs/article/view/732

Bruce, Thomas R., and Diane Hillmann. 2004. "The Continuum of Metadata Quality: Defining, Expressing, Exploiting." In *Metadata in Practice*, eds. Diane I. Hillmann and Elaine L. Westbrooks (Chicago: ALA Editions)

Chapman, John W., David Reynolds, and Sarah A. Shreeves. 2009. "Repository Metadata: Approaches and Challenges." *Cataloging & Classification Quarterly* 47 (3-4): 309-325. https://doi.org/10.1080/01639370902735020

Downey, Moira. 2019. "Assessing Author Identifiers: Preparing for a Linked Data Approach to Name Authority Control in an Institutional Repository Context." *Journal of Library Metadata* 19 (1-2): 117-136. https://doi.org/10.1080/19386389.2019.1590936

Gross, Tina, Arlene G. Taylor, and Daniel N. Joudrey. 2015. "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching." *Cataloging & Classification Quarterly* 53 (1): 1-39. https://doi.org/10.1080/01639374.2014.917447

Karen, Calhoun. 2006. The Changing Nature of the Catalog and its Integration with Other Discovery Tools: Final report. Library of Congress. Accessed 22 November 2021. http://www.loc.gov/catdir/calhoun-report-final.pdf

Konstantinou, Nikolaos, Dimitrios-Emmanuel Spanos, Nikos Houssos, and Nikolaos Mitrou. 2014. "Exposing scholarly information as Linked Open Data: RDFizing DSpace contents." *The Electronic Library* 32 (6): 834-851. https://doi.org/10.1108/EL-12-2012-0156

Kumar, Vinit. 2018. "A Model for Content Enrichment of Institutional Repositories Using Linked Data." *Journal of Web Librarianship* 12 (1): 46-62. https://doi.org/10.1080/19322909.2017.1392271

Lubas, Rebecca L. 2009. "Defining Best Practices in Electronic Thesis and Dissertation Metadata." *Journal of Library Metadata* 9 (3-4): 252-263. https://doi.org/10.1080/19386380903405165

Mering, Margaret. 2019. "Transforming the Quality of Metadata in Institutional Repositories." *The Serials Librarian* 76 (1-4): 79-82. https://doi.org/10.1080/0361526X.2019.1540270

Moulaison Sandy, Heather, and Felicity Dykas. 2016. "High-Quality Metadata and Repository Staffing: Perceptions of United States–Based OpenDOAR Participants." *Cataloging & Classification Quarterly* 54 (2): 101-116. https://doi.org/10.1080/01639374.2015.1116480

JLIS.it

Niso Framework Working Group. 2007. A Framework of Guidance for Building Good Digital Collections. 3rd edition. National Information Standards Organization (NISO). http://www.niso.org/publications/rp/framework3.pdf

Park, Jung-Ran. 2009. "Metadata Quality in Digital Repositories: A Survey of the Current State of the Art." *Cataloging & Classification Quarterly* 47 (3-4):213-228. https://doi.org/10.1080/01639370902737240

Park, Jung-Ran, and Yuji Tosaka. 2010. "Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms." *Cataloging & Classification Quarterly* 48 (8): 696-715. https://doi.org/10.1080/01639374.2010.508711

Powell, Andy, Michael Day, and Marieke Guy. 2004. "Improving the Quality of Metadata in Eprint Archives." *Ariadne* (38). Accessed 22 November 2021. http://www.ariadne.ac.uk/issue/38/guy/

Radio, Erik. 2014. "Information Continuity: A Temporal Approach to Assessing Metadata and Organizational Quality in an Institutional Repository." In *Metadata and Semantics Research*, edited by Sissi Closs, Rudi Studer, Emmanouel Garoufallou and Miguel-Angel Sicilia, 226-237. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-13674-5_22

Smith-Yoshimura, Karen, Micah Altman, Michael Conlon, Ana Lupe Cristán, Laura Dawson, Joanne Dunham, Thom Hickey, Daniel Hook, Wolfram Horstmann, Andrew MacEwan, Philip Schreur, Laura Smart, Melanie Wacker, Saskia Woutersen, and Oclc Research. 2014. *Registering Researchers in Authority Files*. Accessed 22 November 2021. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014.pdf

Solomou, Georgia, and Dimitrios Koutsomitropoulos. 2015. "Towards an evaluation of semantic searching in digital repositories: a DSpace case-study." *Program* 49 (1): 63-90. https://doi.org/10.1108/PROG-07-2013-0037

Stein, Ayla, Kelly J. Applegate, and Seth Robbins. 2017. "Achieving and Maintaining Metadata Quality: Toward a Sustainable Workflow for the IDEALS Institutional Repository." *Cataloging & Classification Quarterly* 55 (7-8): 644-666. https://doi.org/10.1080/01639374.2017.1358786

Svantesson, Lotta, and Monica Steletti. 2019. "DSpace ORCID integration: name authority control solution at the European University Institute." Presented at the The 14th International Conference on Open Repositories (OR2019), Hamburg, Germany https://doi.org/10.5281/ZENODO.3553926

Swan, Alma, and Leslie Carr. 2008. "Institutions, Their Repositories and the Web." *Serials Review* 34 (1): 31-35. https://doi.org/10.1016/j.serrev.2007.12.006

Thomas, R. Bruce, and Hillmann Diane. 2004. "The Continuum of Metadata Quality: Defining, Expressing, Exploiting." In *Metadata in Practice*. Chicago: ALA editions.

White, H. C., S. Chen, and G. Liu. 2018. "Relationships between metadata application and downloads in an institutional repository of an American law school." *LIBRES* 28 (1): 13-24. https://www.libres-ejournal.info/2608/

Zhu, Lihong. 2019. "The Future of Authority Control: Issues and Trends in the Linked Data Environment." *Journal of Library Metadata* 19 (3-4): 215-238. https://doi.org/10.1080/19386389.2019.1688368

In the mangrove society: a collaborative Legal Deposit management hypothesis for the preservation of and permanent access to the national cultural heritage*

Giuliano Genetasio[a], Elda Merenda[b], Chiara Storti[c]

a) Biblioteca Nazionale Centrale di Roma, http://orcid.org/0000-0002-6764-4850
b) Biblioteca Nazionale Centrale di Roma, https://orcid.org/0000-0002-5727-8690
c) Biblioteca Nazionale Centrale di Firenze

ABSTRACT

Legal deposit, regulated by Law no. 106 of 15 April 2004 and Presidential Decree no. 252 of 3 May 2006, requires Italian publishers to deposit a copy of the published material with several libraries. Legal deposit involves long-term preservation and access to information on various media, not least computer networks. While traditional media are well regulated, digital legal deposit rules are barely sketched out. The National Central Library of Florence (BNCF), with the National Central Library of Rome (BNCR) and the Marciana National Library of Venice, created Magazzini digitali: a digital legal deposit project that allows harvesting of doctoral theses and e-journals produced by research institutions, in addition to ebooks and commercial journals. Thanks to a collaboration with Horizons and Giunti, BNCR has started an experimental deposit of ebooks through MLOL. While awaiting the regulation on digital legal deposit, it is urgent to reopen the debate on this issue and make more effective the collaboration between institutions involved in the management of the digital library heritage, so as to establish a coordination structure that will define the scientific guidelines and the appropriate technological and service choices.

KEYWORDS

Legal deposit governance; Digital legal deposit; National archive of Italian publishing production; National Central Library of Rome; National Central Library of Florence; Magazzini digitali.

JLIS.it

Legal Deposit and bibliographic control in the digital ecosystem

The National Central Library of Rome (BNCR) and the National Central Library of Florence (BNCF) are responsible for the collection, preservation, and cataloguing of Italian publishing production, under the Regulations for State public libraries (Presidential Decree 5 July 1995, no. 417, art. 1, para. 2). This task is possible primarily because of national laws on Legal Deposit (Law 15 April 2004, no. 106) and previous similar legislation in place in many Italian pre-unitary States. The Regulation on Legal Deposit currently in force (Presidential Decree 3 May 2006, no. 252) establishes that all documents of cultural interest which are intended for public use and produced in whole or in part in Italy must be delivered to the two National Central Libraries, for the creation of a National Archive of Italian publishing production, and to a certain number of libraries in the territory to which the publisher or the person responsible for the publication belongs, for the creation of a Regional Archive of Italian publishing production.[1] Legal Deposit constitutes, therefore, the main acquisition channel of publications for the two National Central Libraries. It is instrumental to the bibliographical control and constitutes its necessary premise. For this reason, the control of publishing compliance carried out by the two National Central Libraries, and still aimed only at publications distributed on traditional media, is of fundamental importance. It is interesting to reflect on why even in the digital ecosystem Legal Deposit continues to play an important role. We are immersed in an endless availability of databases, repertories and bibliographical indexes of different nature and origin, which often allow access, under certain conditions, not only to the bibliographical record but to the full resource. An ecosystem in which crowdsourcing appears to be the only way to truly govern the enormous mass of data and information produced by contemporary society. It is precisely these peculiarities of the digital information ecosystem that make Legal Deposit even more important: to preserve and make permanently accessible the national cultural heritage, a public authority must be identified to govern the entire process of acquiring, managing, and making available documents of any nature and on any medium.

Italian and European legislation

The regulatory framework of Legal Deposit in Italy is completed by the Ministerial Decrees of December no. 28, 2007 and December 10, 2009, which identify the beneficiary institutions of regional Legal Deposit, as well as further agreements, notes, and clarifications by the General Directorate for Libraries and Copyright on specific aspects of Legal Deposit. However, the legislation has several grey areas. First, there is the problem of a lack of coordination among depository institutions. Despite the existence of a Commission for Legal Deposit, which should meet periodically to control and monitor the implementation of the law and to define single issues related to it, and even though the General Directorate for Libraries has clarified some doubtful points, there is no coordinating body among depository institutions for the daily activities related to Legal Deposit, either at the national or regional level. There is also a lack of shared databases to monitor publisher activity and compliance. The regulations currently in force present sever-

---

[1] For previous legislation see (Alloatti 2008, 25–33).

JLIS.it

al critical points linked to the growing phenomenon of paper self-publications:[2] among them, the problem of identifying the subject obliged to make the Legal Deposit stands out, as well as that of partial exemptions, based on the criterion of circulation (publications printed in less than 200 copies are not to be sent to the BNCR). The circulation criterion is difficult to verify both for conventional publishing and self-publications because the circulation is seldom stated in the publication. Finally, another critical point is the total exemptions, which still do not include, even under certain conditions, the exclusion of self-publications. Another problematic aspect is the number of copies for paper Legal Deposit, as many as four between National and Regional Archives of Italian Publishing Production,[3] reduced to three with Decree-Law April 24, 2014, no. 66, for many regions. Even three copies are too many if they involve the multiplication of identical tasks across multiple libraries. Although the legislation provides that some particular types of documents must be deposited in specific institutions designated to manage them, it does not yet apply to all kinds of materials. A further critical point is the monitoring of the fulfillment of the Legal Deposit (Presidential Decree 3 May 2006, no. 252, art. 41), which the Regulation entrusts to generic control instruments of the depository institutions. BNCF and BNCR have databases that cross-reference OPAC and trade book data based on ISBNs and can allow for sufficiently effective monitoring of publishing compliance. At the regional level, however, there do not appear to be similarly adequate monitoring tools. Regional Legal Deposit is an innovative regulatory element because it introduces a partial decentralization of Legal Deposit, but it is also a critical element because the identification of depository libraries and the criteria for dividing the material to be deposited has been difficult. Regional Legal Deposit has not always been successful.[4] Equally thorny is the question of penalties for failure to comply with the Legal Deposit requirement. In the face of widespread evasion, the system of sanctions set out in the legislation involves lengthy and cumbersome procedures.[5] A comparison of documents received for Legal Deposit in 2019 at BNCR and BNCF leaves no doubt. For BNCR: 43205 monographs, 3410 children's publications, 3517 school texts, 387 sheet music and scores, 175 maps, 2329 audiovisuals (CDs, DVDs, and multimedia); the final tally of minor publications is yet to be made. For BNCF: 64184 monographs, 1790 children's, 3311 scholastic, 852 sheet music and musical scores, 123 maps, 2007 audiovisual, 2971 minor publications. These differences show the degree of evasion in both institutions, although partly due to partial exemptions and different categorization criteria. Furthermore, this penalty system tends to criminalize the publisher and thereby make him an enemy, rather than an ally, of the library, with harmful consequences for both parties. Finally, one of the weakest points of the legislation is that of the digital Legal Deposit (Presidential Decree 3 May 2006, no. 252, art. 37), the definition of which is postponed to a further regulation to be formulated – as yet non-existent but soon to be published. Article 37 of the Regulations refers to "voluntary forms of experimentation" of Digital Legal Deposit, through agreements with publishers. Despite the agreement between the Italian

---

[2] A survey conducted in 2015 by the AIE (Italian Publishers Association) had shown how even then almost 50% of ebooks published in Italy were self-publishing, for a total of over 25,000 titles per year. See («Quasi un titolo ebook su due è nato con il self publishing. L'indagine Aie: grandi numeri, ma il mercato è meno di un terzo» 2016).

[3] Presidential Decree 3 May 2006, no. 252, art. 1, para. 2 and art. 6.

[4] See (AIB Biblioteche e servizi nazionali 2020).

[5] Law 15 April 2004, no. 106, art. 7; Presidential Decree 3 May 2006, no. 252, art. 11, art. 43, 44, and 45.

JLIS.it

Ministry of Cultural Heritage and publishers, and despite *Magazzini digitali* (see below), the Digital Legal Deposit still lacks a regulatory framework that goes beyond the experimental phase. It is no longer possible considering the increasingly important Digital Legal Deposit as the younger brother of the paper Legal Deposit. Today it would be appropriate to overcome this obsolete vision and, on the contrary, to speak no longer of digital (or paper) Legal Deposit but of Legal Deposit *tout court*, since the two aspects are increasingly intertwined. Together with the enactment of the Regulation on the deposit of documents distributed via computer networks and the partial reformulation of Chapter VII of Presidential Decree 3 May 2006, no. 252, the Italian legislator has another opportunity to provide libraries with the appropriate tools to carry out their tasks: the planned stage of transposition into national law of the EU *Directive 2019/790 on Copyright in the Digital Single Market*,[6] approved in February 2019. The Directive has received different opinions, sometimes conflicting, from professionals and associations of libraries and librarians[7] but, certainly, the way in which it will be transposed may influence in one way or another the ability to preserve digital memory and especially the possibility of creating services on the digital documentation deposited and preserved: think of e-lending, but also of the cataloguing and indexing of deposited resources through text and data mining activities. The sustainability of Legal Deposit and bibliographic control services in the near future will depend, to a large extent, on how national and European legislators will integrate the point of view of memory institutions – libraries, archives, and museums – on preservation services and permanent access to information, within the reference legislation, both the one specific to Legal Deposit and the one related to copyright, privacy, and personal data processing.

Preserving digital memory: the experience of Magazzini digitali and the challenges of today

On December 15, 2020, the digital archive of *La Stampa*, one of the most important and long-lived Italian national newspapers, disappeared from the web.[8] The archive was built in 2010 with Flash, a technology that, even then, though widespread, was considered risky, mainly because it was proprietary to Adobe, and therefore incompatible with the paradigms of the web, whose optimal functioning is guaranteed by the use of open standards. At the end of 2020, the archives of *La Stampa* had to "close for maintenance" until February 15th 2021 when, after a re-engineering of the platform, it is available again. In the meantime, because the digitization of the journal had been released under a Creative Commons (CC-BY-NC-ND-it 2.5) license, a team at the Internet Archive downloaded the entire archive and made it available by creating a special collection.[9] We can safely say that that of *La Stampa* is by no means an isolated case. For different reasons, in 2017,

---

[6] See https://eur-lex.europa.eu/eli/dir/2019/790/oj.

[7] See («Le raccomandazioni della rete MAB per il recepimento della direttiva europea sul copyright» 2020). See also («Copyright reform (archived) - European Bureau of Library Information and Documentation Associations (EBLIDA)» s.d.).

[8] (Tedeschini-Lalli 2021) and («L'archivio storico de La Stampa sarà di nuovo consultabile entro metà febbraio 2021» 2020).

[9] *Archivio storico La Stampa*, "Internet Archive", https://archive.org/details/la-stampa-newspaper (last accessed March 31, 2021).

JLIS.it

the digital archive of *L'Unità* met the same end.[10] We could cite the cases of thousands, or even millions – in the case of websites[11] – of digital resources of extreme interest that have vanished into thin air.[12] The preservation of digital resources and their accessibility in the long term is a de facto necessity of contemporary society, a responsibility that memory institutions cannot shirk. For this reason, despite the absence of regulations in the field of Legal Deposit of publications diffused through computer networks, the prototype experience of *Magazzini Digitali*,[13] born as such on the basis of Presidential Decree 3 May 2006, no. 252, art. 37, has been consolidated over the years to become a real service. As of December 2020, *Magazzini Digitali* includes:

- over 170,000 doctoral dissertations, through harvesting with OAI-PMH protocol from the repositories of 58 Italian universities;
- about 180 open access journals, through harvesting with OAI-PMH protocol from the repositories of 7 Italian universities and other research organizations or associations;
- about 500 ebooks deposited in BagIt format;
- about 80 TB of high-resolution digital copies from library digitization projects (with Google Books Project as the main source).[14]

Besides, beginning in 2018,[15] BNCF launched a Web archiving service joined by approximately 220 institutions for a total of about 300 sites subject to periodic archiving, ensuring long-term access to digital resources is an organizational, management, and technological challenge, even more than preserving them. In 2020 the Web archiving service made important steps:

- Bibliographic records related to doctoral dissertations and open access e-journals have been indexed in the catalogs (OPAC) of BNCF and BNCR, and archived resources made available by the internal networks of the two Institutes;
- the collections of websites archived by BNCF have been made publicly accessible on the Archive-it platform.

BNCR has an experimental service that focuses precisely on user utilization. BNCR and Horizons Unlimited LLC (the Italian company that creates MLOL – MediaLibraryOnLine, an Italian e-lending platform) reached an agreement in 2019 to start experimenting with free ebook deposit, which was initially joined by Giunti publishers and will be joined by Mondadori publishers during 2021. The pilot project foresaw an innovative infrastructure through a new MLOL Reader Desktop App, which uses for the first time in Italy the Readium LCP DRM, the open-source DRM system developed by EDRLab (the European branch of the Readium Foundation and member of

---

[10] After a pirated copy of the archive was made available on the deep web for a few months, the archive came back online in October 2018, but it is still unknown to this day who and how made it possible to restore the service. See («Riappare online (con un curioso sottotitolo) l'archivio storico de L'Unità. Mistero sull'autore dell'operazione» 2018). *L'Unità* can be consulted at: *L'Unità*, "Internet Archive", https://archive.org/details/lunita_newspaper (last accessed April 1st, 2021)

[11] See also: https://www.internetlivestats.com/total-number-of-websites/ (last accessed 16/12/2020).

[12] See (Laakso, Matthias, e Jahn 2021)the responsibility rested primarily with librarians, but the shift toward digital publishing and, in particular, the introduction of open access (OA.

[13] (Bergamin e Messina 2010, 144–53).

[14] https://www.bncf.firenze.sbn.it/biblioteca/magazzini-digitali/. See also (Storti 2019).

[15] The BNCF carried out an experimental harvesting session in 2006 in which 7 terabytes of data were collected from websites belonging to the .it domain. These data are available through the Archive-it BNCF Collection: https://archive-it.org/home/BNCF. See also (Bergamin 2012, 170–74).

JLIS.it

the W3C for the maintenance of EPUB)[16]. BNCR immediately welcomed the chance of experimenting with a service for the deposit and use of digital content that allows users to have immediate access to documents, without having to wait for the often very long processing times of paper documents.[17] The possible future developments of accessory search and access services (the full text of the documents, the data mining procedures on the books, the possibility of enrichment of the OPAC dialogue with search engines, etc.) are also worthy of interest.

From deposit as a procedure to deposit as a process: a new design for managing complexity

Giovanni Bergamin summarized in 2006 the different positions following the enactment of Law no. 106 of 15 April 2004, which for the first time extended the obligation of Legal Deposit to documents circulated via computer networks: "impossible, useless, civil."[18] Public opinion has had to reconsider the usefulness of such an operation, recognizing that the preservation of digital memory is a duty for any "civilized" society. Instead, there is a debate about the possibility of carrying out this activity in an effective (concerning the available technologies) and efficient (concerning the amount of digital information currently being produced and growing exponentially) way. In addition to the lack of a clear regulatory framework, this possibility still seems remote because there has been too much focus on technology and too little on process management, as has happened in other areas involved in digital transformation. In the words of Luciano Floridi, who inspired the title of this contribution, "the challenge is not technological innovation but digital governance."[19] This governance cannot be implemented by applying or, worse, bending the paradigms of the pre-digital and pre-web world to the web world, but requires a "new design." What is needed is a model that can manage the entire information flow, i.e., the process that enables long-term preservation and access to information, not just individual digital information or a single repository process. It is also an opportunity to reiterate that there is no longer a clear separation between an analog and a digital information flow, as already mentioned. As Luciano Floridi said, we live in a mangrove society[20] in which there is no longer a solution of continuity between offline and online.[21] The new governance model should consider the management of Legal Deposit regardless of the nature of the objects or the form by which information is recorded and transmitted: […] "The term deposit should not be understood in a literal sense (physically bringing an object to a particular location) but in the context of "Legal Deposit" as an "institution" where operation-

---

[16] See https://www.edrlab.org/readium-lcp/ and (Rosenblatt 2017).

[17] MLOL acted as an intermediary between the Library and the partner publishers, setting up procedures for collecting data and bibliographic metadata of the ebooks, and packaging them in a special deposit package. The BNCR user who accesses the Desktop MLOL Reader App from local workstations can download locally the ebooks equipped with the new Readium LCP copy protection system.

[18] (Bergamin 2012).

[19] The quotations by Luciano Floridi are taken from the conference "For an ethics of technology" - Cubò, Bologna, February 13, 2020. See also (Floridi 2017, 2020).

[20] (Floridi 2018).

[21] See (Floridi 2015).

JLIS.it

al procedures must take into account the nature of the object".[22] To design this new governance model, it is necessary to attempt to briefly reflect on the main individual aspects that make up the Legal Depository process, and that have been affected by the digital transformation.

New roles and new actors

Under the law, responsibility for managing the institution of Legal Deposit, including long-term preservation and access to digital resources, is essentially shared between the State and the Regions. The State exercises it through the National Central Libraries of Florence and Rome, the Regions through the institutes identified by the Ministerial Decrees of 2007 and 2009. The traditional management of Legal Deposit allowed co-responsible institutions to proceed independently, often implementing practices that were supplementary in means and ends. The advent of digital technology makes this form of management impractical or inadvisable. A digital copy can be stored, cataloged, and indexed only once and still be accessible in geographically distant points. Moreover, the traditional management conflicts with other regulatory provisions, particularly the provisions of AGID (Agenzia per l'Italia Digitale) relating to the rationalization of public information and communication technology assets.[23] Until now, preservation of and access to resources has been almost exclusively the responsibility of depository institutions. Today, however, it is necessary to involve producers, distributors, and those responsible for information and records in any capacity, depository institutions, and organizations.[24] This does not mean that depository institutions lose or delegate part of their tasks, but that they rediscover a new role and actively bring new players into the resource management process. It has recently been estimated that the entire digital universe consists of approximately 44 zettabytes of data.[25] Even limiting the responsibility of memory institutions "to documents [of cultural interest] intended for public use […] produced totally or partially in Italy",[26] the mass of information to be managed is enormous and constantly growing. Unlike traditional media, thinking especially of websites, it is increasingly difficult to establish cultural interest from the origin: libraries will play a leading role precisely in the ability to define what is of cultural interest, a role that is not new to the traditional responsibilities and tasks of a library but unprecedented for the quantity and types of material to be selected. Digital and the web have not only changed a large part of the traditional publishing industry and market

---

[22] (Bergamin 2012).

[23] «Razionalizzazione del Patrimonio ICT|Agenzia per l'Italia digitale». Accessed April 1st, 2021. https://www.agid.gov.it/it/infrastrutture/razionalizzazione-del-patrimonio-ict.

[24] The need for close collaboration between publishers or producers of information and depository institutions was already evident in 2011, during the phase of definition of the parameters of the experimentation of the Digital Legal Deposit service. On July 14 the General Directorate for Libraries and the most representative associations of the publishing industry signed an agreement "for the promotion of the convention for the legal deposit of digital documents and the license for their use". Although the agreement did not have the expected operational results, it constitutes an important model for the collaborative management of digital legal deposit. See also General Directorate for Libraries website: Direzione generale Biblioteche e diritto d'autore. «Accordo MiBAC - Associazioni Editori». Accessed April 1st, 2021. https://www.librari.beniculturali.it/it/notizie/notizia/4ee4df59-4819-11e1-88f7-b7fd06d12128/.

[25] Tremolada, Luca. «Quanti dati sono generati in un giorno?» Info Data (blog - Il Sole 24 ore), 14 maggio 2019. Accessed April 1st, 2021. https://www.infodata.ilsole24ore.com/2019/05/14/quanti-dati-sono-generati-in-un-giorno/.

[26] Law 15 April, 2004, no. 106, art. 1, para. 1 and 3.

– think of the explosion of the self-publishing and print-on-demand phenomenon – but has also introduced new media or new ways of using existing content: social media, streaming video, podcast platforms, apps, etc. Depository institutions are thus no longer just conservators but selectors of resources. Digital preservation isn't a process that begins the moment the document enters the library's collections: the archivability[27] of a website and, in full, of a digital document must be an original "property" of the documents. The functions of the depository institutions should be reconsidered: the new task will be to define criteria and guidelines for the archivability of documents, taking into account, the theoretical models of reference and the state of the art of technologies on one hand and the real possibilities for information producers to adapt to these models or to adopt standards on the other hand. Similarly, cataloging and indexing can no longer be the exclusive prerogative of libraries, which activate the service as the resources enter the collections. Document producers or distributors should be involved. However, the definition of ontologies for the descriptive, semantic, and managerial metadata remains the responsibility of the depository institutions. Long-term access largely depends on the correct compilation of metadata that describes the policies for access, use, and reuse of documents:[28] cataloging and ordering of resources remains, even in the digital world, the first form of guarantee of access to these resources. Full-text indexing and refinement of search algorithms help facilitate information retrieval, but the number of documents and information that must be retrieved is, as repeatedly stated, increasing. It is therefore clear that the management of Legal Deposit, as a process with the characteristics previously identified, requires, in the first instance, a reorganization of personnel and workflows within the depository institutions. Legal Deposit is a completely new activity in some respects, while in others it continues some of the established services of libraries, which should, however, be revised in light of the changing ecosystem in which it operates. To provide an effective and efficient national service for the preservation and access of digital resources, aligned with international best practices (and indeed capable of constituting a model in its own right), and in keeping with current modes of cultural production, depository institutions should have a sufficient number of professionals capable of fully managing these activities in an integrated manner, not only from a scientific and technological point of view but also from an administrative and organizational one.

New models for acquiring, storing, and accessing resources

Legal Deposit as an institution becomes more understandable in the resource acquisition phase: the sending by the producer or person responsible for the digital publication is considered as a residual modality for the digital deposit, while automatic or semi-automatic harvesting of computer networks becomes the norm. As for the acquisition of resources, memory institutions should define

---

[27] "Archivability" refers to the set of characteristics that the content, structure, functionalities, and interfaces of a site should possess for the site to be preserved and made accessible over the long term with current web archiving tools. The concept is however extendable to other digital resources whether they are spread on the web or not. See also Biblioteca Nazionale Centrale di Firenze. «Archiviabilità dei siti web». Accessed April 1st, 2021. https://www.bncf.firenze.sbn.it/biblioteca/archiviabilita-dei-siti-web/.

[28] This system should provide for the possibility of managing any changes in rights over time, both those provided for by law (think of the term after which a work falls into the public domain) and for cases in which special or advance licenses are issued.

JLIS.it

clear models[29] that always take into account the sustainability of procedures for all stakeholders and the characteristics of digital documents. In summary, the resource acquisition model should define what resources to acquire and how to acquire them,[30] what formats and what protocols, what descriptive and management metadata are required, what metadata for preservation must be attributed during acquisition.[31] Access models must take into account copyright regulations, policies, and licenses established by the owners of the information, user profiles, indexing systems, and last but not least, the possibilities of replay systems. Finally, storage models should take into account the security and redundancy of data, their growth forecasts, and the ability to move and migrate data and documents as needed. These procedures should be based on globally shared conceptual models, take current technologies into account but not be strictly linked to them, and not disregard the regulatory and application context, here including not least the characteristics of work in and for the public administration.

A hypothesis of collaborative management of the Legal Deposit for the preservation and permanent access to the national cultural heritage

"Unfortunately, even when libraries recognize that they need fresh perspectives, they all too often turn to the academic and library-oriented technical communities that simply reinforce the same problems, rather than widening their reach to the outside world to bring in entirely new ideas and perspectives [...] In the end, libraries have reached an inflection point where they will continue to fade into irrelevance when it comes to web archiving if they are not dragged kicking and screaming out of the third century BC and into the modern world. Such modernization can only come from reaching outside of their traditional confines and engaging in sustained partnerships with the outside technical community, bringing in fresh perspectives and approaches. Until then, our web history continues to rapidly slip away, lost forever".[32]

Let us try to summarize the main peculiarities of the governance model of the national Legal Deposit service that have emerged so far:

- Digital preservation is not just secure backup,[33] so it is not an activity that can be solved with simple storage services;
- The preservation of digital cultural heritage is the task of memory institutions. It differs

---

[29] The theoretical reference model for digital preservation remains the OAIS model - Open Archival Information System, ISO 14721 standard. "Models" mean in this context the level of application procedures, therefore not strictly related to technologies and formats in current use. See (Michetti 2008, 32–49). About the latest revision of the OAIS standard see «OAIS: entro breve la revisione». ParER - Polo archivistico dell'Emilia-Romagna. Accessed April 1st 2021. https://poloar-chivistico.regione.emilia-romagna.it/news-in-evidenza/oais-entro-breve-la-revisione.

[30] "The mere accumulation of data, while waiting for more powerful computers, more sophisticated software, and new human skills, will not work, not least because we do not possess sufficient storage capacity." (Floridi 2017, 18). "In hyperhistory, saving is the default option. The problem becomes what to delete." (Floridi 2017, 22).

[31] The reference standard for the production of preservation metadata is PREMIS (PREservation Metadata: Implementation Strategies), «PREMIS: Preservation Metadata Maintenance Activity (Library of Congress)». Accessed April 1st 2021. http://www.loc.gov/standards/premis/.

[32] (Leetaru 2021).

[33] (Bergamin 2018).

JLIS.it

from "standard preservation"[34] in terms of both the object of the preservation (documents of cultural interest vs. computerized documents) and the objectives of the service (preserving the products of national culture and scientific research vs. guaranteeing the evidentiary value of the documents) while presenting necessary similarities in terms of technological solutions;

- Long term preservation is above all a service that has to do with democracy: access to information with equal opportunities for all citizens can only be guaranteed by public institutions, and cannot be the exclusive prerogative of private companies, even if they have a public purpose;[35]

- Digital management requires a revision and rationalization of the policies of protection, valorization, and access to cultural heritage.

Italian public institutions should find, also at the regulatory level, the definitive confirmation of their role for the fulfillment of the tasks of Legal Deposit, bibliographic control, conservation, and access to documents, in whatever form they are recorded and transmitted. A simple investiture by law, in itself, does not qualify them to exercise this responsibility. In the absence of economic, instrumental, and human resources, no kind of governance is possible: the preservation of and permanent access to collective memory is a strategic national service, a challenge that can only be met if tackled collaboratively. A viable solution to these problems, one which is in line with current laws, is a public company to manage the services of conservation and permanent access to cultural heritage. This society should be shared by the bodies and institutions responsible for Legal Deposit: the State (Ministry of Cultural Heritage and its institutes), the Regions, the libraries of the Constitutional Bodies, which receive the deposit upon request of official publications of the State and other public bodies, and CNR – National Research Council, which receives the deposit of publications in the technical-scientific area. It would be important to establish a greater synergy with the governing bodies of SBN (Italian National Library Service), concerning the managing of digital resources within the SBN catalogue (i.e., the Italian Union Catalog), and AGID. It would be important to involve public players such as the Istituto Poligrafico e Zecca dello Stato, ISTAT (National Institute of Statistics), CRUI (Conference of Rectors of Italian Universities) as representatives of the world of academic research, but also, to mention materials of a different nature, RAI Teche. All these institutions are major producers of cultural information or public sources and bearers of similar instances and needs, as well as domain know-how. On this last point, partnership with private entities is essential and there are many candidates for this role. One possibility is to consolidate long-standing synergies with publishers or digital content distribution platforms: by way of example, Casalini Libri and MLOL. Then, we should turn our attention to companies that deal with the preservation and management of digital information: together with the aforementioned Internet Archive and the Wikimedia movement, whose institutional mission is similar

---

[34] Legislative Decree 7 March 2005, n. 82, art. 34 para. 1 bis, and the Decree of the President of the Council of Ministers 3 December 2013, art. 5 para. 3.

[35] The most important player in this field is certainly Internet Archive: https://archive.org/about/. For the same reason, a close collaboration with the Wikimedia Foundation, and with the national chapter Wikimedia Italy, would be highly desirable, especially on the side of the definition of tools and methods of access to preserved documentation.

JLIS.it

to that of Memory Institutes, it is certainly equally important to finding forms of collaboration with the so-called Big Players active in this field, such as Google and Amazon. The constitution of a public company would respond more easily to the need to pool stable resources to make existing services more efficient, and would allow the procurement of resources, especially human and instrumental ones, in a manner more in keeping with the rapid development of technological services. Currently, the costs amounted to approximately € 500,000 per year. They had been calculated based on existing contracts and the experience gained in almost 10 years of Magazzini Digitali, net, however, of the investments already made over the years, in particular by BNCF, and those related to personnel and infrastructure currently of the depository institutions. There are also the costs of the traditional legal deposit. BNCR, for example, invests about € 265,000 per year in external support for the management of bibliographic control, made necessary by the severe shortage of staff that forces the use of external collaborators to carry out the service. These costs are those necessary to guarantee the minimum services for the management of the Legal deposit. The procurement of such substantial resources can no longer be solely tied to specific projects or targeted funding, as was the case in the early stages of testing and implementation of the service. A public company would have a stand-alone budget, established through the co-partnership of the various entities responsible for Legal Deposit,[36] and might be able to provide a share of commercial long-term preservation services to third parties beyond the provisions of Legal Deposit. A company with a public shareholding could recruit personnel more easily by selecting the professional skills not available in the roles of the public administration, or not available in sufficient quantity, albeit always with public evidence procedures. A company with public shareholding would have a flexible corporate architecture, allowing the acquisition of instrumental resources in the technological sphere, without partial or total interruption of services, and loss of know-how. The synergy between institutions, professionals, different skills is the only way that can guarantee relevant results, averting the failure of traditional library policies envisaged by Kalev Leetaru in the contribution published in 2017 on Forbes.com, with the significant title *Why Are Libraries Failing At Web Archiving And Are We Losing Our Digital History?*

---

[36] Regarding the participation of state institutions, resources should be stable and linked to a specific budget chapter.

JLIS.it

References

AIB Biblioteche e servizi nazionali, Commissione nazionale. 2020. «Il deposito legale regionale in Italia: stato dell'arte e risultati di una recente indagine». *AIB Studi* 59 (3): 423–52. doi:10.2426/aibstudi-12019.

Alloatti, Franca. 2008. «L'attuazione della Legge 106 tra incognite e speranze». *Biblioteche oggi* 26 (1): 25–33.

Bergamin, Giovanni. 2012. «La raccolta dei siti web: un test per il dominio "punto it"». *DigItalia* 2 (0): 170–74.

———. 2018. «Conservazione del patrimonio culturale digitale». In *60° Congresso nazionale AIB, 22-23 novembre 2018.*

Bergamin, Giovanni, e Maurizio Messina. 2010. «Magazzini digitali: dal prototipo al servizio». *DigItalia* 2 (0): 144–53.

Biblioteca Nazionale Centrale di Firenze. «Archiviabilità dei siti web». Accessed April 1st, 2021. https://www.bncf.firenze.sbn.it/biblioteca/archiviabilita-dei-siti-web/.

«Copyright reform (archived) - European Bureau of Library Information and Documentation Associations (EBLIDA)». Accessed October 2, 2021. http://www.eblida.org/about-eblida/archive/copyright-reform/.

Direzione generale Biblioteche e diritto d'autore. «Accordo MiBAC - Associazioni Editori». Accessed April 1st, 2021. https://www.librari.beniculturali.it/it/notizie/notizia/4ee4df59-4819-11e1-88f7-b7fd06d12128/.

Floridi, Luciano, a c. di. 2015. *The Onlife Manifesto: Being Human in a Hyperconnected Era.* Cham: Springer International Publishing. doi:10.1007/978-3-319-04093-6.

———. 2017. *La quarta rivoluzione: Come l'infosfera sta trasformando il mondo.* Milano: Raffaello Cortina Editore.

———. 2018. «The good web. Some challenges and strategies to realise it». In *The Web Conference, Lyon, France 23-27 april 2018.*

———. 2020. *Pensare l'infosfera. La filosofia come design concettuale.* Milano: Raffaello Cortina Editore.

Laakso, Mikael, Lisa Matthias, e Najko Jahn. 2021. «Open Is Not Forever: A Study of Vanished Open Access Journals». *Journal of the Association for Information Science and Technology* 72 (9): 1099–1112. doi:10.1002/asi.24460.

«L'archivio storico de La Stampa sarà di nuovo consultabile entro metà febbraio 2021». 2020. *lastampa.it.* dicembre 14. Accessed October 2, 2021. https://www.lastampa.it/rubriche/public-editor/2020/12/14/news/l-archivio-storico-de-la-stampa-sara-di-nuovo-consultabile-entro-meta-febbraio-2021-1.39659298.

«Le raccomandazioni della rete MAB per il recepimento della direttiva europea sul copyright». 2020. *AIB-WEB.* ottobre 18 Accessed October 2, 2021. https://www.aib.it/attivita/mab/2020/85856-raccomandazioni-mab-recepimento-direttiva-europea-copyright/.

JLIS.it

Leetaru, Kalev. 2021. «Why Are Libraries Failing At Web Archiving And Are We Losing Our Digital History?» *Forbes*. Accessed April 1st, 2021. https://www.forbes.com/sites/kalevleetaru/2017/03/27/why-are-libraries-failing-at-web-archiving-and-are-we-losing-our-digital-history/?sh=f22dacb6ecd4.

Michetti, Giovanni. 2008. «Il modello OAIS». *DigItalia* 1 (0): 32–49.

«OAIS: entro breve la revisione». ParER - Polo archivistico dell'Emilia-Romagna. Accessed April 1st 2021. https://poloarchivistico.regione.emilia-romagna.it/news-in-evidenza/oais-entro-breve-la-revisione.

«PREMIS: Preservation Metadata Maintenance Activity (Library of Congress)». Accessed April 1st 2021. http://www.loc.gov/standards/premis/.

«Quasi un titolo ebook su due è nato con il self publishing. L'indagine Aie: grandi numeri, ma il mercato è meno di un terzo». 2016. *Prima online*. dicembre 10. Accessed October 2, 2021. https://www.primaonline.it/2016/12/10/251136/quasi-un-titolo-ebook-su-due-e-nato-con-il-self-publishing-lindagine-aie-grandi-numeri-ma-il-mercato-e-meno-di-un-terzo/.

«Razionalizzazione del Patrimonio ICT|Agenzia per l'Italia digitale». Accessed April 1st, 2021. https://www.agid.gov.it/it/infrastrutture/razionalizzazione-del-patrimonio-ict.

«Riappare online (con un curioso sottotitolo) l'archivio storico de L'Unità. Mistero sull'autore dell'operazione». 2018. *Prima online*. ottobre 16. Accessed October 2, 2021. https://www.primaonline.it/2018/10/16/279189/riappare-online-con-un-curioso-sottotitolo-larchivio-storico-de-lunita-mistero-sullautore-delloperazione/.

Rosenblatt, Bill. 2017. «Readium LCP Set to Launch». *Copyright and Technology*. marzo 11. Accessed October 2, 2021. https://copyrightandtechnology.com/2017/03/11/readium-lcp-set-to-launch/.

Storti, Chiara. 2019. «Storage, enhancement and preservation of doctoral dissertations in the experience "Magazzini digitali": a contribution to research and access». *JLIS.it* 10 (1): 114–24. doi:10.4403/jlis.it-12526.

Tedeschini-Lalli, Mario. 2021. «Tecnologia Digitale Obsoleta, Un Secolo e Mezzo Di Storia a Rischio». *Medium*. febbraio 23. Accessed October 2, 2021. https://tedeschini.medium.com/tecnologia-digitale-obsoleta-un-secolo-e-mezzo-di-storia-a-rischio-1bb75bf68c2f.

Tremolada, Luca. 2019. «Quanti dati sono generati in un giorno?» Info Data (blog - Il Sole 24 ore), maggio 14. Accessed April 1st, 2021. https://www.infodata.ilsole24ore.com/2019/05/14/quanti-dati-sono-generati-in-un-giorno/.

JLIS.it

Thesauri in the Digital Ecosystem

Anna Lucarelli[(a)]

a) Biblioteca nazionale centrale di Firenze

ABSTRACT

In recent years, thesauri have taken on new roles, new functions, and have shown some advantages over other knowledge organizazion systems (KOS). They are increasingly important in the linked data environment of the semantic web. The *Nuovo soggettario*, created and maintained by the National Central Library of Florence, is an example of the changing uses of controlled subject systems, like thesauri and subject heading lists. Thesauri are shown to be dynamic tools, essential components for the integration of data on the web, especially for mapping and to assist with interoperability among heterogeneous resources. With the adoption of formats of the semantic web, such as RDF/SKOS, and following international standards, thesauri have evolved and have proven to be increasingly useful with free reuse and across various frameworks. To varying degrees, they have enabled increased multilingualism and conceptual equivalences, connecting information and metadata produced by institutions of different countries. As authority control systems, they interact with Wikidata and help build 'bridges' between worlds that were too far apart until not long ago, namely libraries, archives, and museums. Will the challenge of search engines, machine learning and artificial intelligence override the thesauri or will it make them even more involved?

KEYWORDS

JLIS.it

Thesauri and bibliographic control

Since the beginning of IFLA's UBC Programme, universal bibliographic control has been primarily focused on the sharing and standardization of descriptive cataloguing. Talking about thesauri gives rise to the following questions:

- the current state and the possible new future of subject indexing of which thesauri are essential components, along with subject heading lists (Petrucciani 2019, 163-173);
- the path followed by subject indexing in recent years; a path that is considered 'autonomous' compared to other cataloguing processes; a path strongly connected to the procedures and to the languages used in various countries, in several cultural, geographical and, above all, linguistic contexts;
- the relationship of thesauri with other knowledge organization systems (KOS), such as ontologies, classifications, taxonomies, and so on (Gnoli 2020);
- the role they play in current bibliographic control, considering that the concept of bibliographic control in the digital ecosystem is changing and is evolving *Dalla catalogazione alla metadazione* (from cataloguing to creating metadata), just to use the title of a recent book (Guerrini 2020) and "transitioning to the next generation of metadata" (Smith-Yoshimura 2020). This is a period when cataloguing tools, data management, and infrastructures are more than ever crossing transitional borders, tied to strategies to make bibliographic data more visible on the web.[1]

We now have an opportunity for a better integration of subject data in universal bibliographic control.

As we will see, thesauri have taken on new roles, new functions and shown some advantages over other knowledge organization systems (KOS). Yet, there are many arguments and confrontations on this issue.

In their multiple types (general or specialized domains; polyhierarchical or monohierarchical, monolingual or multilingual, etc.), thesauri continue to prove their effectiveness compared to simple lexicons or *flat lists*; they have proved to be versatile, usable, both in the framework of the post-coordinated and pre-coordinated languages, in which the rules for the citation order in the subject strings are added to vocabulary control.

Controlled vocabularies are studied by the Subject Analysis and Access Section of IFLA[2], but even by the International Society for Knowledge Organization (ISKO) with its regional chapters.[3] Thesauri are also handled by terminology associations,[4] a transversal discipline. Unfortunately, the communities that are involved are not always interactive among one another. The relationship between terminology experts and librarians engaged in subject indexing still continues to be weak and not as creative as it could be.

---

[1] For a selective bibliography on the current state of the subject indexing, specifically with regard to the French reality, see: *L'indexation matière en transition: de la réforme de Rameau à l'indexation automatique* 2020.

[2] https://www.ifla.org/subject-analysis-and-access.

[3] https://www.isko.org/.

[4] E.g., Associazione italiana per la terminologia (Ass.I.Term): http://www.assiterm91.it/.

JLIS.it

Even thanks to their formalized structures, thesauri have been significantly supported by standardization. Subject indexing is one of the rare library tasks specifically regulated by the International Organization for Standardization (ISO). Let's mention the ISO 5963:1985 standard (*Methods for examining documents, determining their subjects, and selecting indexing terms*) on the conceptual analysis, recently validated in 2020, and ISO 25964:2011-2013 (*Thesauri and interoperability with other vocabularies*), just concerning the thesauri themselves, and renewing ISO-5964:1985 and ISO-2788:1986, established before the digital universe existed. However, these are not the only standards about both documentation and terminology. Within the Italian framework, groups and technical committees of Commissione UNI CT/014[5] also deal with them.

Rise or fall of thesauri?

The national libraries have tried to make their vocabularies, used for subject indexing, more 'visible' and usable, through various modes of integration with their own OPACs or with the open data hubs.

Subject indexing in libraries has however suffered a slowdown, not only because the data referred to semantic contents is considered to be less necessary, but also because indexing is a labor-intensive and hence expensive process. The lack of human resources is not only an Italian problem, though. Nevertheless, in recent years, the development of thesauri has spread everywhere.

This spread is studied by BARTOC, a database developed and maintained at the University of Basel, since 2013. It inventories various systems for the organization of knowledge, including web applications and mapping, so far totalling 3,393 (data as of November 2021).[6]

In particular for thesauri, BARTOC monitors the establishment of thesauri across the world, by describing them by their main features, identifying 781 thesauri to date. Almost all of them are in digital format and freely accessible on the web by open licenses, many of them being multilingual. Thesauri are inventoried and assessed by librarians and institutes around the world. Assessment includes for their performance, for their compliance with the standards, and for their level of semantic coverage.

They are assessed according to their ability to handle their own terminology expansion, starting with a particular *corpora*, and for their ability to be representative with regard to specialized domains (Folino and Parisi 2020). Such issues in Italy are examined not only by librarians, but also by CNR Institutes and by centres of excellence such as the Laboratorio di Documentazione dell'Università della Calabria.[7]

Thesauri are further assessed for the possibility of being integrated with algorithms employed for the automated indexing.

Ultimately, they are assessed according to the quality of their data (e.g., ability to be re-used) and according to the sustainability of the resulting costs, also considering that the personnel involved in creating and maintaining thesauri are required an indispensable professional development.

---

[5] https://www.uni.com/index.php?option=com_uniot&view=struct&id=853557&Itemid=2447.

[6] http://bartoc.org/.

[7] https://www.labdoc.it/.

JLIS.it

We can think that the development of these tools may depend on the fact that the terminology has acquired more and more importance within "metadata-ing". Yet, it is not only a matter of this. No longer limited to the library and documentation world, these tools have actually gone beyond the context of subject indexing and of information retrieval (IR); they have been involved in other 'universes'.

Following the standardization established by ISO 25964 and by RDF/SKOS,[8] thesauri organize both the concepts and the terms by which they are represented. The central role of the concept (that is a unit of thought, rather than a lexical element of a specific language), has ensured that the borders which separate thesauri from other knowledge organization systems have become more fluid. There are two reasons for this: for the possibility of the correct *reconciliation* of expressions in different languages, and for the opportunity to compare different systems starting from the conceptual cores on which they are established. It is no coincidence that "metathesauri" have also been set up (e.g., UMLS by the National Library of Medicine in the United States[9]).

An approaching process has been activated among classifications, schemes based on subject headings and ontologies; in the latter, the relationships among concepts are less standardized. The very relational structure of thesauri, when rigorous, encourages their evolution towards the ontologies (Biagetti 2018; Biagetti 2020).

Vanda Broughton, Leonard Will and Stella Dextre Clarke have faced such interesting issues in a recent series of virtual classes organized in 2020-2021 by ISKO UK.[10]

One characteristic of thesauri is that of being dynamic tools, obviously linked to the linguistic fabric of the context in which they are established, yet, often, through multilingual functionalities. In addition, thesauri have shown their capabilities, not only as 'tools of the trade' for librarians but even as tools for users. Yet, we know that this happens if they are provided in the right way, if they are 'well integrated' in OPACs, and if librarians employ them also as a support to reference service and to information literacy (Ballestra 2011, 395-401).

The reason why they are so costly is due to the constant maintenance work they require, along with a careful supervision of the increase mechanisms in relation to the 'literary warrant'.

They also require a continuous assessment of their structural coherence, a monitoring of the semantic relationships, particularly for synonymy and lexical variants on the one side, polysemy and new meanings on the other side.

Languages (of works, of users, of catalogues) quickly evolve, so the work to be carried out on neologisms is continuous. To give a current example, let's think about the importance to 'control' concepts connected to the pandemic we are living in and which are the subjects of works already published. SARS-CoV-2; COVID-19; Social distancing; Confinement; Lockdown; Contact tracing... (see Figures 1-5). These terms were added promptly and captured some of these new concepts from the first months of 2020. Not all vocabularies acquire new concepts with the same timeliness.[11]

---

[8] https://www.w3.org/2004/02/skos/.

[9] https://www.nlm.nih.gov/research/umls/index.html.

[10] https://www.iskouk.org/KOED.

[11] For a first survey on the terms tied to COVID-19 pandemic, inserted into Thesaurus of *Nuovo soggettario* already since March 2020: Francioni and Lucarelli 2020. On the Italian words about pandemic also Accademia della Crusca: https://accademiadellacrusca.it/it/contenuti/lacruscaacasa-le-parole-della-pandemia/7945.

JLIS.it

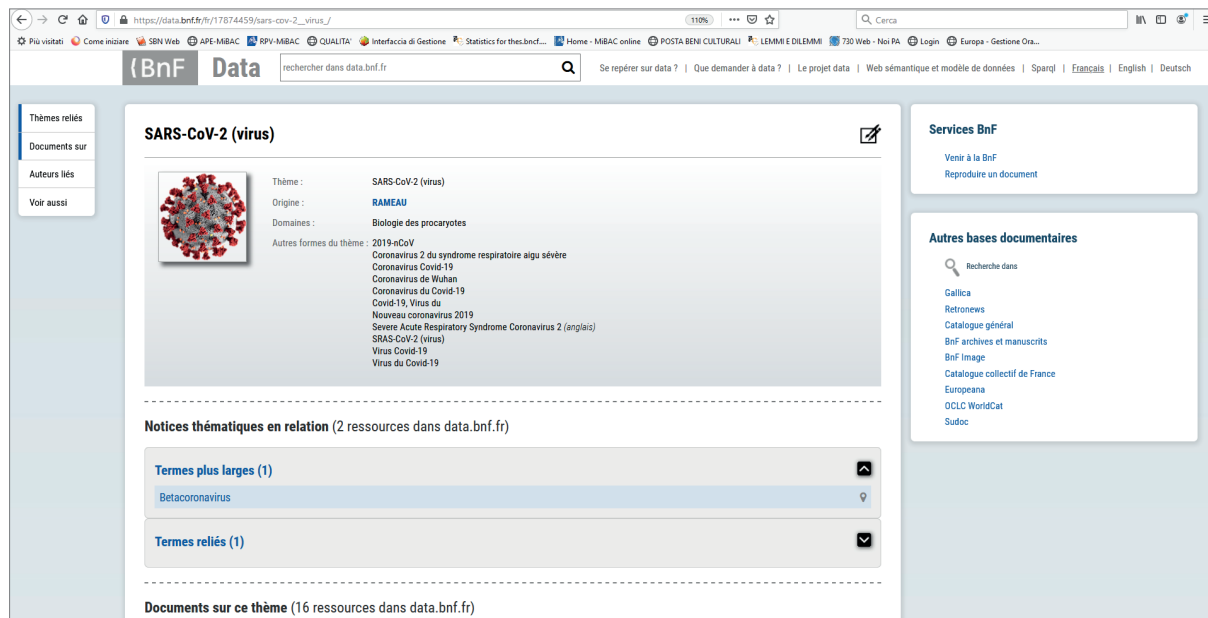Examples of new concepts related to the current world health crisis:
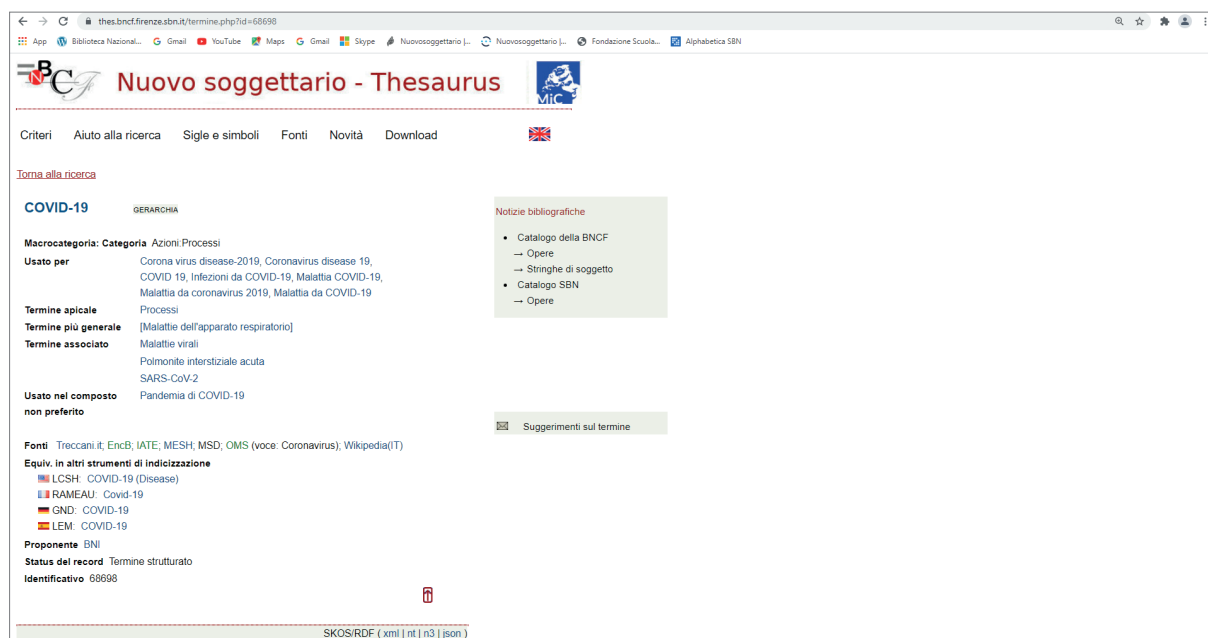

Fig. 1. The concept *SARS-CoV-2 (virus)* in RAMEAU

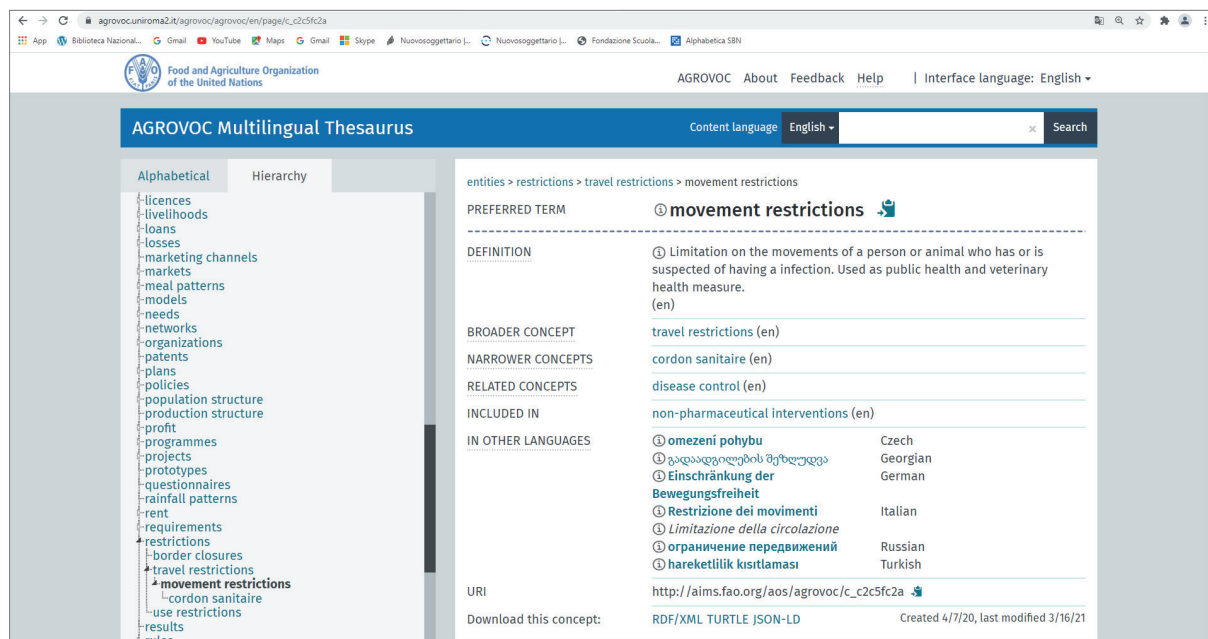Fig. 2. The concept *COVID-19* in *Nuovo soggettario*

JLIS.it
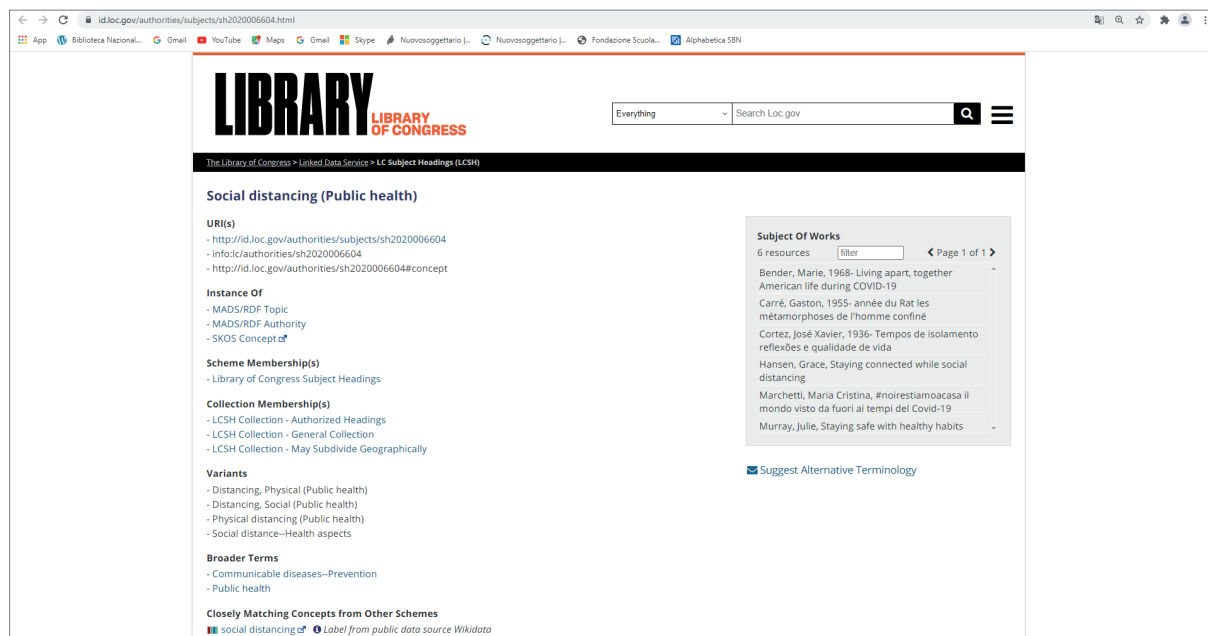

Fig. 3. The concept *Movement restrictions* in AGROVOC
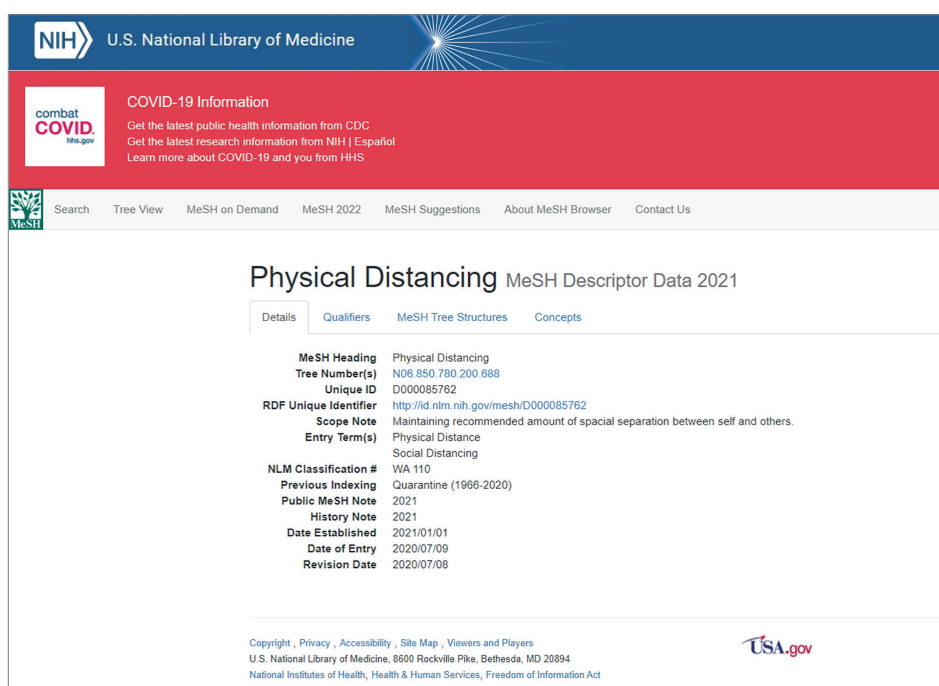

Fig. 4. The concept *Social distancing (Public health)* in LCSH

JLIS.it

Fig. 5. The concept *Physical Distancing* in MESH

Integration of data on the web

What is important is that thesauri have proved to be the essential components for the integration of data on the web and thus fundamental elements for the affirmation of the semantic web.

We are dealing with the role of the thesauri within the semantic web at various levels and in various contexts and there are many studies on this topic (e.g., Martinez-González and Alvite Díez 2019). We could say that they are among 'the best friends' of the semantic web, for their capability to provide metadata in RDF, that is to say in open formats which allow their re-use in the most varied contexts (not necessarily library ones), because they encourage the development of mapping as well as the interoperability between heterogeneous resources (Zeng 2019, 122-146).

When we wonder which is the most re-used data among those processed by libraries, thesauri are a good example.

Many of them are connected with DbPedia.[12] Tens of other thesauri have recently connected to Wikidata.[13] The Italian Thesaurus of *Nuovo soggettario*[14] – created and maintained by the National Central Library of Florence (BNCF) – has had links with Wikipedia since 2007. Since

---

[12] https://wiki.dbpedia.org/.

[13] https://www.wikidata.org/w/index.php?title=Special:WhatLinksHere/Q89560413&limit=500.

[14] https://thes.bncf.firenze.sbn.it/ricerca.php. BNCF, with an almost centuries-old tradition for subject indexing (started in 1925), has the institutional task to curate the Italian subject indexing tools. *Nuovo soggettario* contains the concepts/terms employed in the framework of a pre-coordinated language that contemplates also the rules on the construction of the subject strings. Yet, thesaurus is obviously usable also for the post-coordinated indexing. It is employed by the Italian National Bibliography (BNI) and by most libraries of the Italy's National Library Service (SBN). It was also presented during IFLA General Conference 2009 (Cheti, Alberto, Anna Lucarelli, and Federica Paradisi. 2009).

JLIS.it

2013, reverse mapping has been implemented with a mutual browsing mechanism as well as with a synchronization realized through the field P508 (BNCF Thesaurus ID) of Wikidata (Lucarelli 2014).[15]
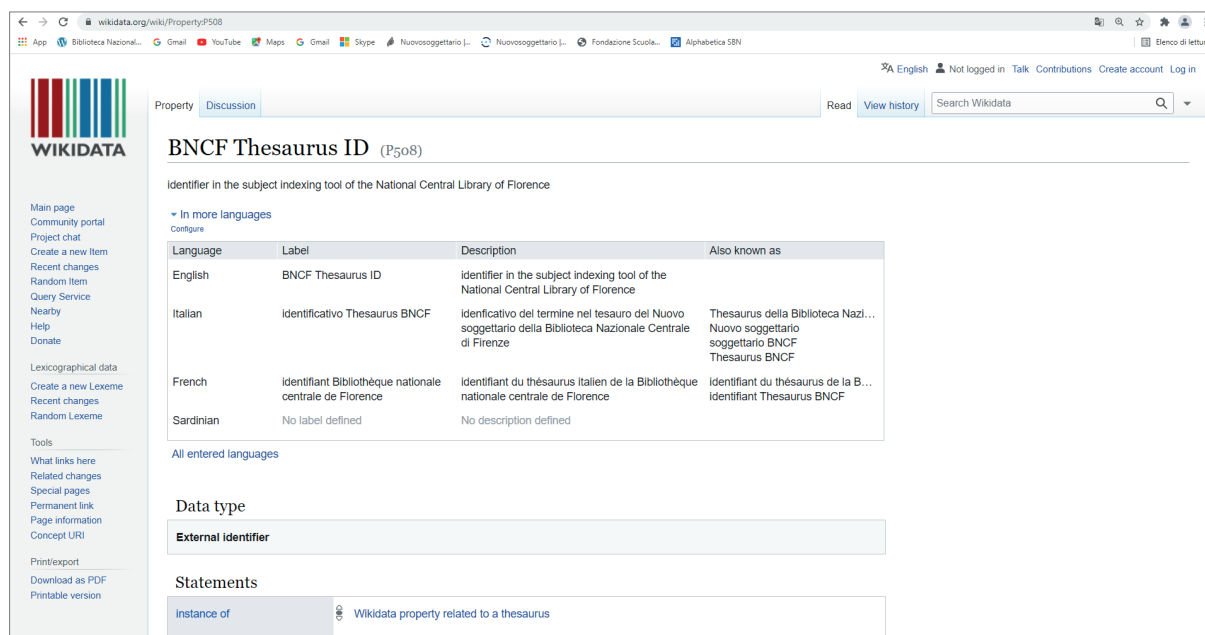
Fig. 6. Thesaurus of the National Central Library of Florence and Wikidata

Since thesauri are among the 'main actors', the interpreters of the semantic web, we must evaluate their costs on the basis of the benefits they bring to the linked open data and on the possibility of creating mapping, as Stella Dextre Clarke has recently reminded us in the above-mentioned virtual classes[16].

The opportunities offered by open data and by mapping, in both the research world and public administrations, are unquestionable. Some examples?

A few years ago, the City of Florence made use of the open data of BNCF's Thesaurus in order to organize the City's open data.

When, in May 2013, the reverse mapping from the Wikipedia entries to the corresponding terms of *Nuovo soggettario* was implemented, the number of visitors to the OPAC of BNCF has increased 28% in only one month.

As we can see from this example, it is no longer possible to talk about thesauri without talking about interoperability. The international standard ISO 25964 dedicates the second of its two parts to the methods for the realization of this interoperability. Furthermore, the interoperability ac-

---

[15] The *Nuovo soggettario* was the first general thesaurus to activate a form of interlinking with a version of Wikipedia in a specific language, preceded, at an international level, by experiences in specialized sectors as in the case of Thesaurus for Economics of Leibniz-Informationszentrum Wirtschaft (http://zbw.eu/stw/version/latest/about).

[16] About the costs resulted from the procedures of the bibliographic control, also Bergamin 2020, p. 167.

JLIS.it

tivated by thesauri has made them fundamental 'hubs' as well as 'bridges' for the connection between data from different institutions. Mapping also has been realized with particular success in the context of multilingualism. Not only multilingual vocabularies (such as the notable AGROVOC, AAT, EUROVOC, IATE, that are often cited, in a crossed mode, interconnecting one another), but also monolingual vocabularies with equivalences in other languages in the form authorized by those other vocabularies or subject heading schemes. At the same time, Pat Riva explained the importance of multilingualism and of the internationalization of the bibliographic description in order to facilitate access (Riva 2021).

In the revolution of open data, thesauri are thus on the 'front line'. Many of them have implemented new formats for the publications and the exchange of metadata (i.e., SKOS) by exceeding the previous ones (i.e., Zthes). They have become structures "of" the web.
In the linked open data cloud, many controlled vocabularies are represented, including those created and maintained by the national libraries.[17]
The Thesaurus of *Nuovo soggettario* has been in SKOS since 2010 and has achieved the 'five stars' of Tim Berners-Lee.[18] It can also be found in the hub *dati.beniculturali* of the Ministero della Cultura.[19]

The initiatives of national libraries and national bibliographies

Since the publication of IFLA's *Guidelines for subject access in National Bibliographies* ten years have passed, but many indicated best practices are still valid.[20] Following these guidelines, both national libraries and national bibliographies that are assigned to the bibliographic control of our countries, have implemented important choices in the field of subject indexing.
Many national libraries have updated their bibliographic tools to follow the latest standards and entered the world wide web of data, following new 'conceptual models'.
For some of these institutions it has been a period of reforms, like for the Bibliothèque Nationale de France which, in 2019, made public its *Réforme de Rameau.*[21]
Regardless of the subject indexing language used, the national libraries continue to benefit from the controlled vocabularies even when indexing graphic resources, audio resources, ancient works and, in certain countries, works of fiction as well. They even use controlled vocabularies when providing Genre/Form descriptions, a practice that is also supported by IFLA.[22] In some cases, they use expressly dedicated thesauri, for the indexing of particular types of resources, for instance, the *Library of Congress Genre/Form Terms for Library and Archival Materials* (LC-GFT).[23]

---

[17] https://lod-cloud.net/clouds/publications-lod.svg.
[18] https://lod-cloud.net/dataset/bncf-ns.
[19] https://dati.beniculturali.it/altri-linked-open-data-del-mibact/.
[20] https://www.ifla.org/publications/ifla-series-on-bibliographic-control-45.
[21] https://rameau.bnf.fr/syntaxe.
[22] https://www.ifla.org/node/8526.
[23] https://id.loc.gov/authorities/genreForms.html.

JLIS.it

National libraries generally use these vocabularies for projects of automated indexing or semi-automated indexing of online resources, by having them interact with implemented algorithms. For example, this is part of the subject cataloguing policies of the Deutsche Nationalbibliothek, as explained by Ulrike Junger since the beginning (Junger 2018), and also more recently described by Mödden and Suominen (Mödden 2021; Suominen 2021).

In the name of the data quality, the use of vocabularies continues to rely on uncontrolled keywords. Thanks to mapping to RDF and to open data's hubs, the national libraries' vocabularies encourage a connection among different OPACs, which hopefully is a prelude to additional future forms of connections; some of these connections were originated from the project named MACS (Multilingual Access to Subjects) which was exceptionally innovative and whose operational phase started in 2005.[24]

As we can see in the figures below, starting a search with the subject term employed by the Deutsche Nationalbibliothek, one sees the connected publications, but it is also possible to move to the French equivalence of data.bnf, where resources on that topic can be explored. Through the correspondent RAMEAU page, it is possible to browse towards *Library of Congress Subject Headings* (LCSH) where works about the same topic can be explored in the catalogue of the Library of Congress. The equivalences are generally ensured by the form of the closely matching concepts from other schemes, as well evidenced by LCSH.

Fig. 7. Concept with equivalences in Gemeinsame Normdatei (GND) and links

---

[24] https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/9041.

JLIS.it
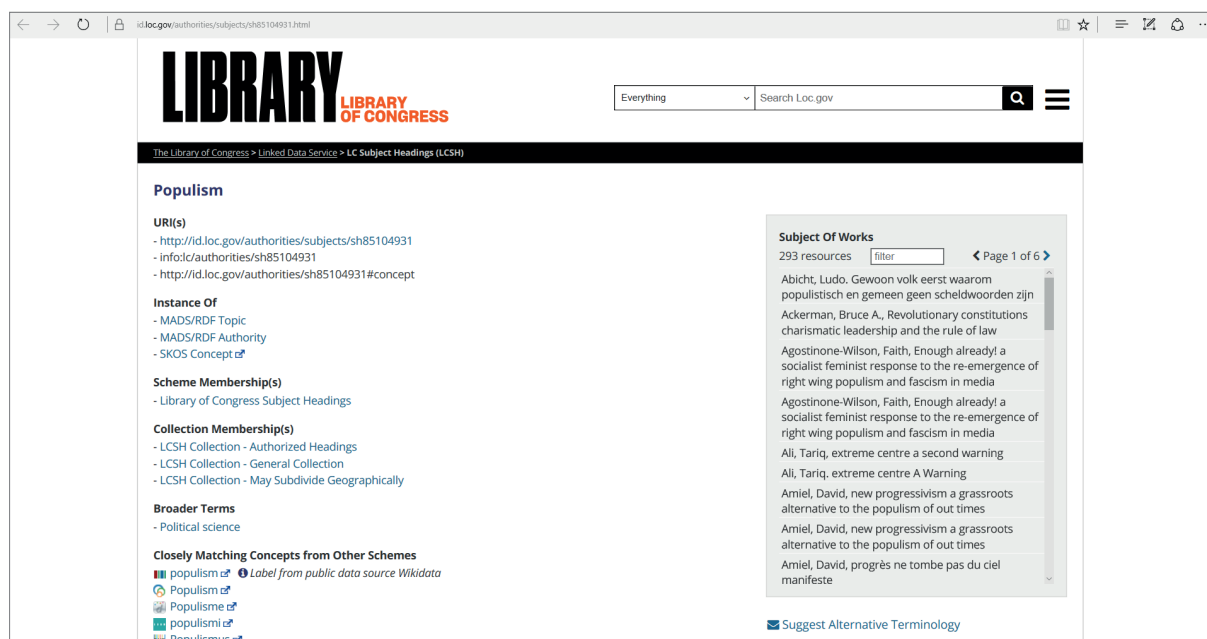
Fig. 8. Concept with equivalences in *Library of Congress Subject Headings* (LCSH) and links

Likewise the Thesaurus of *Nuovo soggettario* has been connected to the works described in the online catalogues of the National Central Library of Florence and Italy's Servizio Bibliotecario Nazionale (SBN), as shown in Figure 9, it has also been possible to navigate to Datos. BNE, that is, to the controlled equivalents of the Biblioteca Nacional de España, and, from there, it has been possible to explore the *Obras* on the same subject in the BNE catalogue.[25]

---

[25] https://datos.bne.es/tema/XX525409.html.
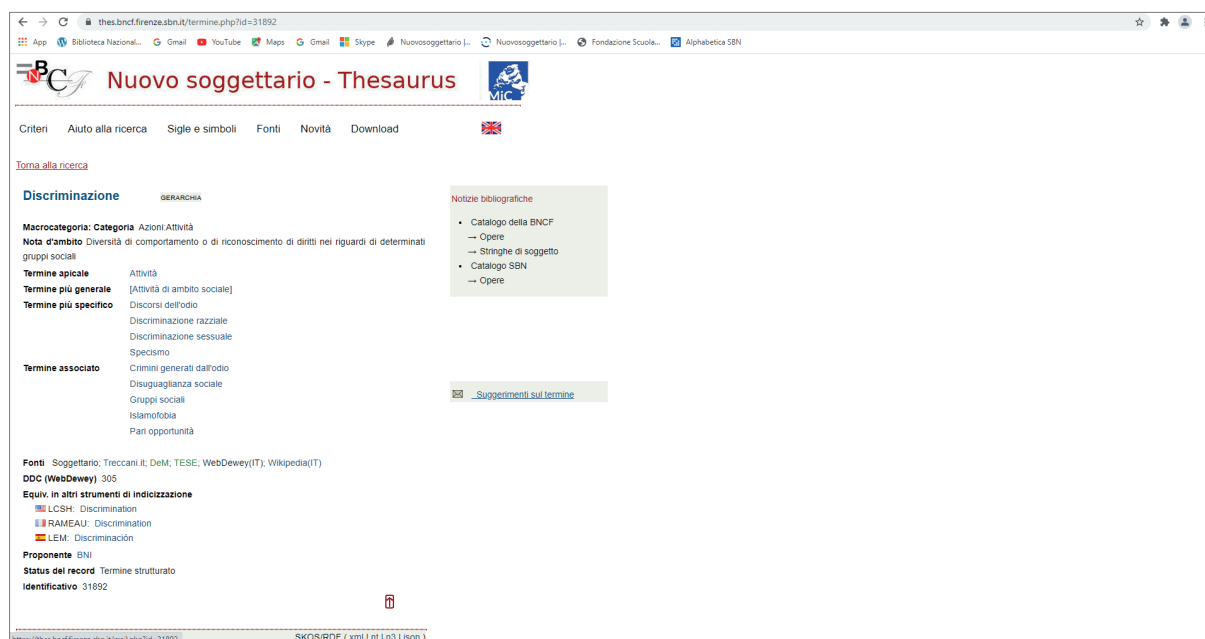
Fig. 9. Concept with equivalences in *Nuovo soggettario* and links[26]
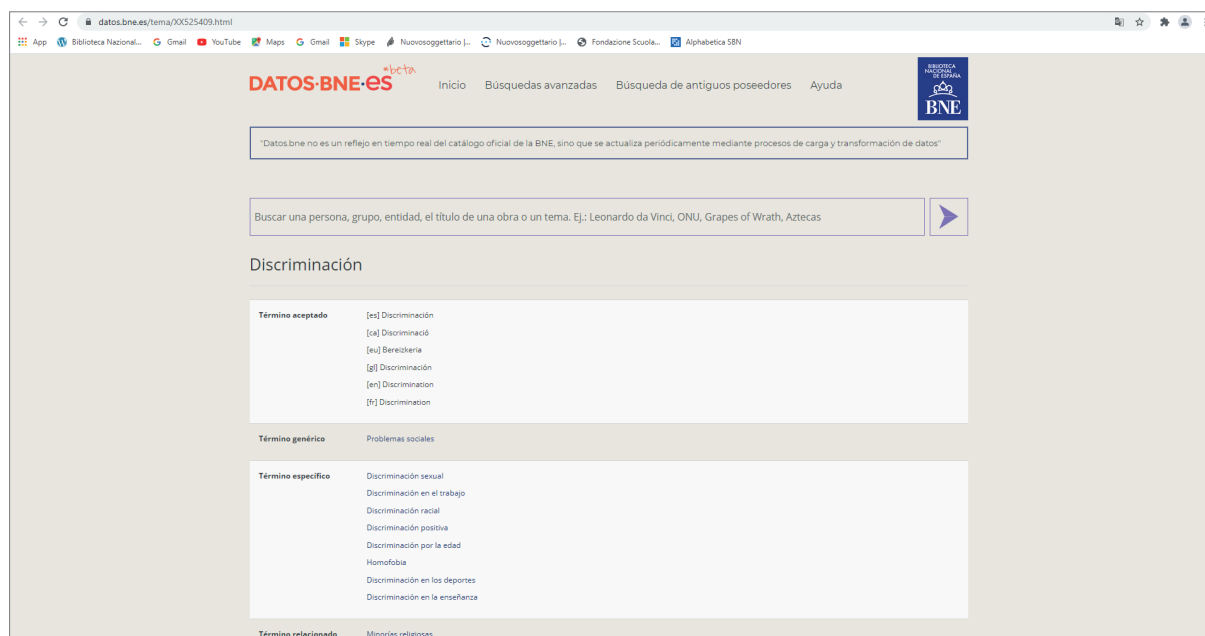
Fig. 10. Concept with equivalences in EMBNE and links

---

[26] The figure refers to the result of the research carried out on the Thesaurus at the time of the Conference (8th-12th February 2021) . For current results, see: https://thes.bncf.firenze.sbn.it/termine.php?id=31892

In fact, over the years, the Italian Thesaurus of *Nuovo soggettario* has considerably increased the mapping with other KOS and with equivalents of other vocabularies and continues to link to more equivalences (Viti 2017, 624-637). The number of links with LCSH has increased from 390 in 2011 to the current 14,970; with the French terms of RAMEAU from 380 in 2012 to the current 13,380 links; with the German terms from 130 in 2018 to the current 2,200; with the Spanish terms from 300 in 2019 to the current 2,270.[27]

Making such links is challenging work, requiring careful mapping and not without problems. For instance, there are challenges about the level of equivalences among concepts, especially across languages. This was explained by Pino Buizza in one of his latest papers on mapping between the Thesaurus of *Nuovo soggettario*, in Italian, and the two subject heading lists produced by national bibliographic agencies in the United States and in France: the *Library of Congress Subject Headings*, in English, and the *Repertoire d'autorité-matière encyclopédique et alphabétique unifié*, in French (Buizza 2020, [59]-68).

The equivalences found in *Nuovo soggettario*, when downloaded through SKOS, can activate mutual connections or give rise to the indication of the variant in Italian, as data.bnf shows in Figure 11 under the term *Épidémies*, among the *Autres formes du thème.*

Such initiatives demonstrate the importance that policies be activated among national libraries.
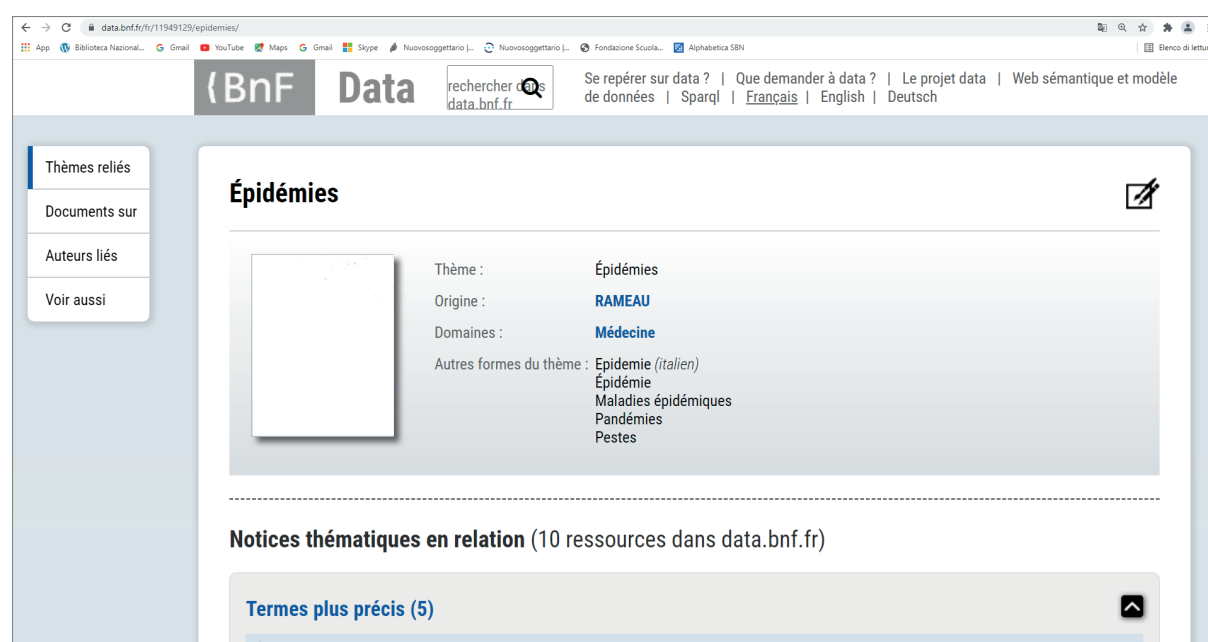
Fig. 11. Mutual connections or the indexing of the variant in Italian

---

[27] The comprehensive data on the trend of the equivalences in other languages are visible in: https://thes.bncf.firenze.sbn.it/stat.php.

JLIS.it

Thesauri and Authority control: connection with other interlocutors

Collaborating on policies does not mean that the different indexing languages used by national libraries and connected through the respective vocabularies must have the same characteristics, the same syntactic rules. Not all tools have the same compliance with the standards, the same structure or functionality. Not all of them are polyhierarchical. Not all of them have comprehensive hierarchies up to the top term. Not all receive both common and proper names. Not all have the same integration with an OPAC and open data.

What brings them together is the progressive alignment among one another, the fact that they achieve common features, for example, to be integrated within Wikidata, so they are all visible on Wikipedia.

Who would have imagined that an encyclopedia would connect its own entries with the most important controlled vocabularies created by national libraries for the purpose of the bibliographic control? At the bottom of the Wikipedia page, you can find the box 'Authority control' with the relevant links.[28]

We know that libraries are not the only producers of bibliographic data, and that other operators are involved in universal bibliographic control. Yet, the data produced by 'certain' major libraries keep reflecting the highest level of quality.

Thesauri and recent features in today's context

Other issues related to thesauri within the digital ecosystem might be added to the above-described panorama. I take a cue from the *Nuovo soggettario* to outline some particularly interesting ones:

1. It has grown in size.
   Nuovo soggettario, in compliance with ISO 25964, has so far had a remarkable quantifiable increase: from 13,000 terms of the prototype to the current 67,000 terms.

2. It has a new interface.
   Since 2020, it has had a new, more user-friendly interface, implemented during the development of BNCF's new web site.

3. It interfaces with classification systems.
   Beyond the above-mentioned multilingualism, *Nuovo soggettario* maps with the Italian WebDewey (Crociani, Giunti, and Viti 2016), etc.

4. It has increased coverage in various subject domains.
   Thanks to the institutions that collaborate with BNCF,[29] it has largely enhanced its general coverage and expanded coverage in specific domains.

---

[28] See, for instance, the connections to the main thesauri at the footnotes of *Arredo urbano* of the Wikipedia in Italian language through the "Controllo di autorità": https://it.wikipedia.org/wiki/Arredo_urbano.

[29] https://thes.bncf.firenze.sbn.it/enti.htm.

5. It can be employed for the Genre/Form indexing.
    This will be possible once our OPACs implement the tag MARC 655.

Also, the Thesaurus of *Nuovo soggettario* is employed apart from BNCF for the subject indexing of specialized resources:

- for audio and audiovisual resources, as for instance, in projects on oral sources of the Istituto centrale per i beni sonori e audiovisivi (ICBSA) (Magrini 2021);
- for graphic resources, for instance for photographs, also in BNCF but additionally in photographic libraries, for example, in the Fototeca - Biblioteca Panizzi;[30] for iconographic resources and maps, for instance in the Museo Galileo (Pocci 2020);[31]
- for archival resources, for example, for the documents indexed in projects of BNCF in collaboration with both Soprintendenza archivistica e bibliografica della Toscana,[32] and Historical Archives of the European Union.[33]

Integration of the *Nuovo soggettario* with databases of archives and museums

This connection of the Thesaurus of *Nuovo soggettario* with databases of archives and museums is quite interesting.

Let's look first at the Gallerie degli Uffizi, one of the most important museums in the world.[34] In 2019, BNCF started a partnership, a "Research pact," with the Uffizi.[35]

From *Violini* of *Nuovo soggettario* [36] it is possible to browse through the records of the Gallerie degli Uffizi catalogue thanks to the connection with *Violino* of the "Scheda OA" (Opere/oggetti d'arte) for the object's definition.[37] A reverse connection can also be seen from the record of the Museum. When the concept from the *Nuovo soggettario* indicates an iconographic subject (for instance *Albero della vita* [tree of life]), the link is to the Uffizi works that represent that subject, as shown in Figure 12.

From some terms, for example *Sestanti* [Sextant] (as shown in Figure 13), it is possible to view the resources of both the Gallerie degli Uffizi and the Museo Galileo.[38]

---

[30] http://panizzi.comune.re.it/Sezione.jsp?titolo=Fototeca&idSezione=233.

[31] https://www.museogalileo.it/it/biblioteca-e-istituto-di-ricerca/biblioteca-digitale/collezioni-tematiche/747-bibliotheca-perspectivae.html.

[32] http://sa-toscana.beniculturali.it/index.php?id=2.

[33] https://www.eui.eu/en/academic-units/historical-archives-of-the-european-union.

[34] https://www.uffizi.it/.

[35] https://www.beniculturali.it/comunicato/uffizi-e-biblioteca-nazionale-di-firenze-patto-per-la-ricerca.

[36] https://thes.bncf.firenze.sbn.it/termine.php?id=17664.

[37] http://www.iccd.beniculturali.it/it/ricercanormative/29/oa-opere-oggetti-d-arte-3_00.

[38] https://thes.bncf.firenze.sbn.it/termine.php?id=30976; http://catalogo.uffizi.it/it/29/ricerca/iccd/?search=*&fromRA=true&filter_OGTD-words=%3D&filter_OGTD=Sestante; https://catalogo.museogalileo.it/oggetto/Sestante.html.

Fig. 12. The Uffizi works on *Albero della vita*

Fig. 13. The term *Sestanti* as seen in A. the *Nuovo soggettario*, B. the Gallerie degli Uffizi's Schede di catalogo, and C. the Museo Galileo's Museo virtuale

JLIS.it

An example of links with archives can be seen in Figure 14, where *Federalisti europei* in the *Nuovo soggettario* is linked with the Ernesto Rossi fund of the Historical Archives of the European Union.[39]
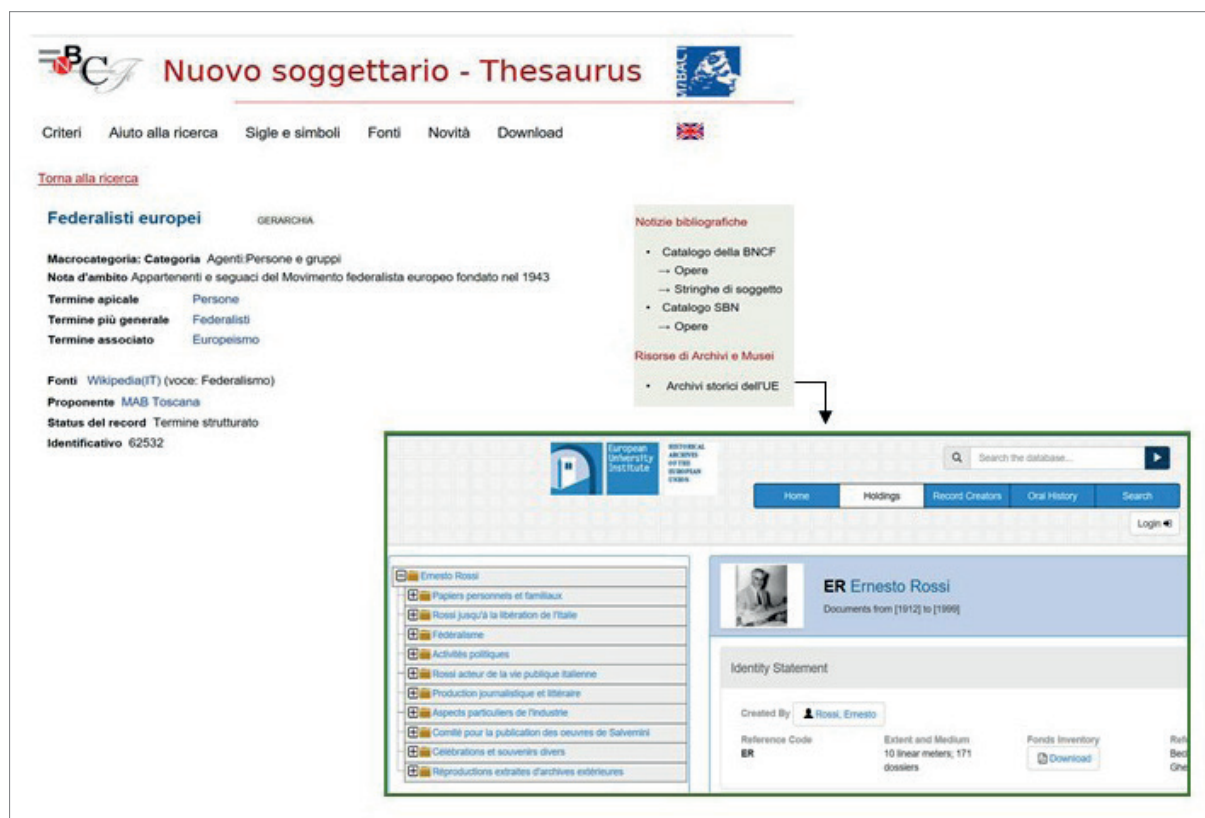

Fig. 14. Links between *Nuovo soggettario* and the Historical Archives of the European Union

These examples of the GLAM (Galleries, Libraries, Archives and Museums) perspective are also promoted by the Wikipedia universe,[40] and in Italy by the MAB (Musei Archivi Biblioteche) projects in which BNCF has participated while joining various research efforts.[41] Likewise, it is hoped there will be future possible connections, for example, with the controlled vocabularies of the Sistema Archivistico Nazionale (SAN)[42] or collaborations with institutions dealing with the standardization of the terminology employed for the cataloguing of the cultural heritage, such as the Istituto Centrale per il Catalogo e la Documentazione (ICCD) (Birrozzi et al. 2020).

---

[39] https://thes.bncf.firenze.sbn.it/termine.php?id=62532; https://archives.eui.eu/en/fonds/115005?item=ER. About the experimentation on subject indexing of Ernesto Rossi Fund: Becherucci et al. 2019, 24-48.

[40] For example, see the recent Gruppo Wikidata per Musei, Archivi e Biblioteche, https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche.

[41] https://www.aib.it/attivita/mab-italia/.

[42] http://san.beniculturali.it/web/san/home;jsessionid=66BD6878BF6E20807ACABB005C45C7CE.sanapp01_portal.

JLIS.it

The perspectives of machine learning, artificial intelligence, automated subject indexing

In which directions will the future of the *Nuovo soggettario* go? Its challenges are not that different from those of other thesauri.

Within the current context, which has much changed due to the predominant role of the Internet, subject indexing is interacting with the semantic capabilities of search engines, such as Google, with the development of both artificial intelligence and machine learning and, of course, with the dissemination of the ever increasing number of digital resources.

At the same time, we know that it is wrong to assume that sources transmitting information be only those 'hooked' by Google, just as it is wrong to confuse the functions of our catalogues with those of other tools for access to information.

However, as often pointed out by the indexing experts, such as Stella Dextre, one must also be aware that libraries and other institutions, dealing with information retrieval, have much fewer resources to be earmarked for 'manual' subject indexing, that is 'intellectual' indexing, as compared to Google's algorithms.

Despite the presence of search engines and their powerful automatic and semi-automatic indexing, the role of thesauri does not seem to be outdated.

For instance, Birger Hjørland, professor at the Royal School of Library and Information Science of Copenhagen, has very recently questioned about the reasons why the search engines, despite they apply principles of semantic type, do not make knowledge organization (KO) and mapping of the relationships among concepts superfluous at all (Hjørland 2021).

'Human' taxonomists working for Google, support the well-known Google Knowledge Graph, which is connected with DBpedia, Wikidata and with the linked data. This is a project about which very many reservations have been expressed.[43]

The procedures for automated and semi-automated translation/indexing are dealt with within IFLA[44] but also within countless other frameworks; to give some examples, in Italy these procedures are studied at the Istituto di linguistica computazionale di Pisa, at the Universities of Padua and Udine. In 2011, BNCF took its first steps by starting a project for the semi-automated indexing of digital doctoral theses. At that time, we used MAUI and other open source software. Should we have the resources and the possibility to restart this project, we could build on the important experiences of other national libraries, such as the Deutsche Nationalbibliothek or utilize tools like those implemented by National Library of Finland.

Studies will continue on machine learning, knowledge graphs like Google's, *corpora* of terms, and the benefits that thesauri can bring to our users, because not only the artificial intelligence world will benefit from such insights, but also libraries and the national bibliograpies world in their mission for the dissemination of knowledge.

In closing, here are some Keywords for the future of thesauri and for their challenges: creativity, versatility, sharing.

---

[43] https://en.wikipedia.org/wiki/Google_Knowledge_Graph.

[44] Automated subject analysis and access Working Group, https://www.ifla.org/node/92551.

JLIS.it

References

(Last consultation of the websites: 15 July 2021).

Ballestra, Laura. 2011. "Information literacy education in Italian libraries: evidence from an Italian University." Bibliothek Forschung und Praxis 35, no. 3 (December):395-401.

Becherucci, Andrea, Silvia Bruni, Benedetta Calonaci, Emilio Capannelli, Walter Fochesato, Anna Lucarelli, and Sonia Puccetti. 2019. "Libri per gli internati militari italiani durante la Seconda guerra mondiale: un inedito di Ernesto Rossi." Biblioteche oggi 37, (May):4-48. DOI: http://dx.doi.org/10.3302/0392-8586-201904-024-1.

Bergamin, Giovanni. 2020. "Postfazione." In Guerrini, Mauro. 2020. Dalla catalogazione alla metadazione. Tracce di un percorso, 167-168. Roma: Associazione italiana biblioteche.

Biagetti, Maria Teresa. 2018. "A comparative analysis and evaluation of bibliographic ontologies." In Challenges and opportunities for knowledge organization in the digital age. Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto, Portugal, edited by Fernanda Ribeiro, Maria Elisa Cerveira, 501-510. Baden Baden: Ergon.

Biagetti, Maria Teresa. 2020. "Ontologies (as knowledge organization systems)." In ISKO Encyclopedia of Knowledge Organization, edited by Birger Hjørland and Claudio Gnoli. https://www.isko.org/cyclo/ontologies.

Birrozzi, Carlo, Barbara Barbaro, Maria Letizia Mancinelli, Antonella Negri, Elena Plances, and Chiara Veninata. 2020. "Catalogare nel 2020. La digitalizzazione del patrimonio culturale." Aedon. Rivista di arti e diritto on line no. 3. http://www.aedon.mulino.it/archivio/2020/3/birrozzi.htm.

Broughton, Vanda. 2020. "General principles underlying knowledge organization systems (KOS)." In *KO-ED Introduction to Knowledge Organization*. https://www.iskouk.org/event-4025408

Buizza, Pino. 2020. "Thesaurus and heading lists: equivalence and asymmetry." In Knowledge Organization at the Interface. Proceedings of the Sixteenth International ISKO Conference, 2020 Aalborg, Denmark, herausgegeben von International Society for Knowledge Organization (ISKO), prof. Marianne Lykke, prof. Tanja Svarre, prof. Mette Skov, Daniel Martinez Avila, [59]-68. Baden-Baden: Ergon.

Cheti, Alberto, Anna Lucarelli, and Federica Paradisi. 2009. "Subject indexing in Italy: recent advances and future perspectives.". https://www.ifla.org/past-wlic/2009/200-lucarelli-en.pdf.

Clarke, Stella Dextre. 2020 "How should today's thesaurus earn its keep?." In *KO-ED Introduction to Knowledge Organization*. https://www.iskouk.org/event-4048801

Clarke, Stella Dextre. 2020 "What is a thesaurus? How and Why so?." In *KO-ED Introduction to Knowledge Organization*. https://www.iskouk.org/event-4048800

Crociani, Laura, Maria Chiara Giunti, and Elisabetta Viti. 2016. "Trent'anni di Dewey in Italia: il ruolo della Biblioteca nazionale centrale di Firenze e i nuovi sviluppi sul fronte dell'interoperabilità con altri strumenti di indicizzazione semantica." AIB studi 56, no. 1 (January/April):87-101. DOI: https://doi.org/10.2426/aibstudi-11408.

JLIS.it

Folino, Antonietta and Francesca Parisi. 2020. "Rappresentatività e copertura semantica dei KOS." AIDAinformazioni 38, no. 3/4:93-112.

Francioni, Elisabetta and Anna Lucarelli. 2020. "Nuovi concetti, nuovi termini ai tempi del Coronavirus." Bibelot: notizie dalle biblioteche toscane 26, no. 1 (January/April). https://riviste.aib.it/index.php/bibelot/article/view/12038.

Gnoli, Claudio. 2020. Introduction to Knowledge Organization. London: Facet Publishing.

Guerrini, Mauro. 2020. Dalla catalogazione alla metadatazione. Tracce di un percorso; prefazione di Barbara B. Tillet; postfazione di Giovanni Bergamin. Roma: Associazione italiana biblioteche.

Hjørland, Birger. 2021. "Search engines and Knowledge Organization (or why we still need Knowledge Organization)." KO-ED Theoretical Perspectives. https://www.iskouk.org/event-4058726.

L'indexation matière en transition: de la réforme de Rameau à l'indexation automatique, sous la direction d'Etienne Cavalié. 2020. https://www.bnf.fr/sites/default/files/2020-03/biblio%20indexation%20matiere%2011mars20.pdf.

Junger, Ulrike. 2018. "Automation first – the subject cataloguing policy of the Deutsche Nationalbibliothek." http://library.ifla.org/2213/1/115-junger-en.pdf.

Lucarelli, Anna. 2014. "«Wikipedia loves libraries»: in Italia è un amore corrisposto…" AIB studi 54, no. 2/3 (May/December). DOI: https://doi.org/10.2426/aibstudi-10108.

Magrini, Sabina. 2021. "«Ti racconto in italiano»: management, description and indexing of oral sources. A project by the ICBSA (Istituto Centrale per I Beni Sonori e Audiovisivi)." In Conference BC 2021. Video. https://www.youtube.com/embed/Yo6Vi72E1T4?start=10942&end=12772.

Martínez-González, M. Mercedes, and María Luisa Alvite Díez. 2019. "Thesauri and semantic web: discussion of the evolution of thesauri toward their integration with the semantic web." IEEE Access, 7. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8873649.

Mödden, Elisabeth. 2021. "Artificial intelligence, machine learning and DDC Short Numbers." In Conference BC 2021. Video. https://www.youtube.com/embed/Yo6Vi72E1T4?start=124&end=1523.

Petrucciani, Alberto. 2019. "C'è un futuro per l'indicizzazione?." In Viaggi a bordo di una parola. Scritti sull'indicizzazione semantica in onore di Alberto Cheti, a cura di Anna Lucarelli, Alberto Petrucciani, Elisabetta Viti; presentazione di Rosa Maiello, 163-173. Roma: Associazione italiana biblioteche.

Pocci, Adele. 2020. "Bibliotheca perspectivae: una sperimentazione del Nuovo soggettario nell'ambito specialistico dell'iconografia scientifica." Bibelot: notizie dalle biblioteche toscane 26, no. 3 (September/December). https://riviste.aib.it/index.php/bibelot/article/view/12798.

Riva, Pat. 2021. "The multilingual challenge in bibliographic description and access." In Conference BC 2021. Video. https://www.youtube.com/embed/Yo6Vi72E1T4?start=12826&end=14355.

Smith-Yoshimura, Karen. 2020. Transitioning to the Next Generation of Metadata. Dublin, OH: OCLC Research. https://doi.org/10.25333/rqgd-b343.

JLIS.it

Suominen, Osma. 2021. "Annif and Finto AI: developing and implementing automated subject indexing." In Conference BC 2021. Video. https://www.youtube.com/embed/Yo6Vi-72E1T4?start=1892&end=2953.

Viti, Elisabetta. 2017. "My First Ten Years: Nuovo soggettario growing, development and integration with other Knowledge Organization Systems." Knowledge Organization 44, 8:624-637.

Will, Leonard. 2020. "From concepts to knowledge organization systems." In *KO-ED Introduction to Knowledge Organization*. https://www.iskouk.org/event-4043820

Zeng, Marcia Lei. 2019. "Interoperability." Knowledge Organization 46, 2:122-146. https://doi.org/10.5771/0943-7444-2019-2-122.

JLIS.it

How to build an «Identifiers' policy»: the BnF use case

Vincent Boulet[a]

a) Bibliothèque nationale de France

ABSTRACT

Identifiers are at the crossroads of two interconnected, major evolutions which heavily impact national libraries: the massification of dataflow, redrawing the place libraries occupy within the global and national data ecosystem in a shared environment, and the strategic shift towards entity management underlying behind the new professional practices and standards. Based on the experience and maturation libraries are gaining in this field, the time maybe has come to formalize them and to highlight the impressive strike force libraries could have in a highly competitive landscape. This is the aim the Bibliothèque nationale de France is trying to reach by publishing an identifiers' policy. It comes as the last part of a triptych after the new cataloguing policy (2016, including the indexing policy published in 2017) and the quality policy (2019). This identifiers' policy is intended to clarify why and on what grounds a national library could, more or less, get involved in a given identifier, taking into account the diversity of scope, governance structure and business model of identifiers, be they international (for instance: ISNI, ISSN, ARK) or local (for instance: the BnF proper identifiers). Therefore, the identifiers' policy highlights why it is necessary to use permanent, trustworthy identifiers and to what extent they are helpful in the daily working and quality control processes led by cataloguers. This is why the identifiers' policy is not limited to principles, but has a very concrete dimension, both for internal and external issues.

JLIS.it

Why an "identifiers' policy" now ?

Libraries' web presence now makes them familiar with identifiers and their uses. This presence poses for them several major and well-known challenges. We can summarize them as follows:

1. The adaptation of their system and data model to the requirements of research and "findability" of their resources *in* the Web. This global framework implies and fuels a needed, major shift of the data structuring, from a world where libraries used to standardize records for making them exchangeable into a world where libraries, along with other players, have to structure data for making them sharable. This issue is at the heart of the crucial problematic of the future of the bibliographic control and has many, crucial implications. For instance, the division of the bibliographic world into bibliographic records and authority records is now close to an end. Therefore the emerging international standards go with the flow, be it the IFLA-LRM data model published by IFLA in 2017 or the new version RDA reshaped by the "3-R project" which became the official version of the RDA international cataloguing code last December. Both have the same underlying principle, namely an entity/relations-based overall model. It means, from the authority control point of view, to switch to logic based on entity management.

2. Resources in a digital world are increasingly more agile and more scalable, due to changes in research and uses' practices. Have we to describe serials or articles published on several platforms? Have to describe coherent set of musical works or a given piece of music diffused by various platforms under various formats? That issue has major implications on legal deposit for digital sound, books and movies. This complex reality challenges the new, above-mentioned library models and cataloguing codes, as they have to take into account changing resources which do not necessary feel part of any idealistic pyramidal model. What is recorded now should not be considered as permanent.

3. The data flows are becoming more and more massive, as the metadata accompanying them. This is also a challenge both for the bibliographic control and for the consistency of library databases. It actually raises the question of how applicable cataloguing rules are for the whole data set libraries deal with. Here is the issue of quality control processes and quality policy, because quality processes can be applied differently according to different data sources and subsets. This makes the question of sourcing data crucial, both for data flows reused by libraries, and for data flows libraries disseminate to end-users.

4. The technical and legal opening of datasets and catalogues is one of the points to be considered for having really sharable data. It may also be a political, strategic commitment taken by public administration towards citizens. As far as the legal opening is concerned, it may also put on the table the issue of mentioning the source of the data and keeping it associated with the metadata produced by a given player.

All these challenges are well-known for the future of the bibliographic control and we have to draw consequences from them. The entity management is unthinkable and impossible without any identifier management and identifiers' policy. The shift from labels (different forms of a name for a person for instance) to identifiers provides less ambiguous data and a kind of stability. Identifiers allow access points or labels to be treated as entities being differently usable according to

context and needs : this is the key-principle of the "nomen" entity in IFLA-LRM. This shift also improves interoperability of data in regards with different contexts[1].

Beyond these principles and opportunities, libraries have nowadays to deal with a very scattered landscape, due to the wide variety in nature offered by identifiers they are using or have intention to use. We can distinguish:

- The global identifiers supported by an ISO standard, which ISO signs an agreement with an international agency about. They correspond to a specific business model and global governance, whose libraries are a part of, along with other players, like music and cultural industry or copyright management firms. Libraries take part in a global business and scientific framework deciding on attribution and possible uses of a given identifier, and they can act as basic members, or a registration center for a given community or a specific field. This is, for instance, the case for ISNI (ISO standard 27729:2012), ISSN (ISO standard 3297), and ISAN (ISO standard 15706-2). For instance, BnF hosts the French national ISSN center and has an official, nationalwide responsibility on this identifier. BnF is, furthermore, an ISNI registration agency since 2014 for a specific dataset, corresponding to the scope of the French legal deposit and national bibliography. But BnF has no special responsibility on ISBN.

- The global identifiers which could be assimilated to a de facto standard, or are being engaged in a standardizing process, and which are supported by an users' community. For instance, ARK (*Archival Resource Key*), an identifier created by *California Digital Library* (CDL), intending for identifying all resources, both physical or digital, records from catalogues or even immaterial resources as concepts. ARK is based on some key-principles and on a community of players engaged to maintain them ("naming authorities", being able to attribute ARK to their resources, and "addressing authorities", being able to resolve the identifier in order to give through it access to resources, by applying a policy of permanence). Moreover, the ARK identifiers have an explicit structure, which make them a de facto standard. So, about ARK, BnF respects an engagement framework with an users' community.

- The specific identifiers BnF has itself set up and is maintaining for internal uses and management of its databases, as for instance internal numbers of bibliographic and authority records (for instance: FRBNF identifiers). But external players can reuse them when reusing these records. So, even if these identifiers have been designed for internal uses at the time of catalogues' automatization, they are also de facto external. BnF keeps the complete control on their maintenance.

So, this short review shows that, over time, successive projects and needs, identifiers have been piled up one on another. In the same time, we have been gaining gradually more maturity and more experience on the overall identifiers' issue.

Managing identifiers doesn't fall from the Jabal Musa as Ten Commandments, but is highly de-

---

[1] Gordon Dunsire and Mirna Wilner, "Authority versus authenticity: the shift from labels to identifiers". In: Authority, provenance, authenticity, evidence: selected papers from the conference and school Authority, provenance, authenticity, evidence, Zadar, Croatia, October 2016. Edited by Mirna Willer, Anne J. Gilliland and Marijana Tomić. Zadar : Sveučilište u Zadru, 2018. p. 87-113.

pending on human and IT resources, on transparency in how these identifiers are managed and on what libraries intend to do with them. Nevertheless, the way of dealing with identifiers as a whole, of choosing them, of handling with them must be consistent with best practices given from a global perspective. We have already framework documents for them, endorsed by W3C[2], by IFLA[3] or by other international authoritative bodies. But, the question, for a given library, could be raised from another perspective. From the point of view of a given institution, to what extent using and disseminating identifiers can be helpful for addressing its own role and tasks? What criteria can be used strategically to justify the commitment of the library in one or more identifiers, and, possibly, its non-involvement? Here is the aim of an identifiers' policy.

Identifiers: a commitment story

Using identifiers highly depends on how committed or engaged libraries want to be. We can easily assume an activist aspect for conceiving and implementing policies. In its identifiers' policy, BnF defines the idea of « engagement » as following:

- For a given and explicit dataset, BnF integrates identifiers in its dataflow and in its development policy regarding metadata (for instance : ARK for every resource, and ISNI for "agent" entities). This is why the identifiers' policy is a follow-up of the BnF quality policy. Identifiers are a tool to delineate specific data subset on which a specific quality control can be applied. It is also helpful to automatize some data processing, by helping interconnections of data. For instance, one of the projects BnF ISNI registration agency is developing is to propose alignments between EAN and ISNI so as to help cataloguers to create links between bibliographic and authority records (and, tomorrow, between manifestations, works and agents).
- BnF ensures, through identifiers, persistence of accessibility to its resources, in a broader meaning of the word: physical resources, digital resources (both digital version of physical documents, and natively digital resources), metadata describing and identifying resources. Identifiers ensure how trustworthy resources are identified for end-users.
- BnF builds up for end-users specific services and transactions thanks to identifiers, being based on its status of national bibliographic agency. For example, the BnF ISNI registration agency has built some transactions with the French book supply chain to register and disseminate ISNIs for their authors.
- BnF disseminates identifiers and resources for free, thanks to legal and technical opening. From this regard, the idenfiers' policy is a follow-up of the open data policy BnF has set up as early as 2011 for data.bnf.fr and as 2014 for every resource.

In other words, the identifiers' policy ensures: to have easily disseminated resources, for the broadest communities, to have traceable, linkable, visible and discoverable resources.

This is the reason why the identifiers' policy is at the crossroads of the strategic shift made by BnF

---

[2] Data on the Web Best Practices, W3C recommendation, 31st January 2017 (https://www.w3.org/TR/2017/REC-dwbp-20170131/)

[3] Best Practice for National Bibliographic Agencies in a Digital Age, https://www.ifla.org/FR/node/8786

JLIS.it

under the name of « bibliographic transition »[4], and embodied by several strategic documents : the statement on open data (2014), the cataloguing policy (2017)[5], the indexing policy (2018)[6] and the quality policy (2019)[7]. A global metadata policy is under preparation and should be published this year.

Negotiating tensions

An identifiers' policy has to deal with three major tensions.

The first tension is the relationship between principles and concrete work and data libraries have to handle with. An identifiers' policy should be intended to give a general framework to action and to concrete involvement on identifiers, both internally, by integrating the identifiers management to the concrete dataflows and cataloguers' work, and externally, for end-users. The question is not to add even more practices for a given identifier, but give practices global framework and direction. In other words, an identifiers' policy finds its role somewhere between a statement of principles on the one hand, and concrete practices and using in the other hand.

The second tension regards the relationship between a common policy and the diversity of identifiers, as said above. It means handling with the diversity of identifiers themselves, and the diversity of how libraries can exercise some responsibility on them. Libraries can only use identifiers in their dataflows, without any significant role ; or they can attribute them ; or they can maintain alignments, or they can build up services for third parties, for instance for the library national community, or the book supply chain.

Here are, for instance, the different role BnF exercises, or intends to exercise on identifiers.

BnF role	International ISO identifiers	Identifiers with an international audience	Local identifiers
Attribution or registration responsability	ISSN, ISNI	ARK	FRBNF
Identifiers BnF doesn't attribute, but BnF uses and builds services for the community on.	ISBN	EAN	
Identifiers which BnF develops alignments with		LCSH, MESH, GND, datos.bne.es, VIAF, NOMISNA, Geonames, Agrovoc, Wikidata	
Identifiers integrated in dataflows	ISAN	EIDR	

---

[4] For more details on the « Bibliographic transition » national programme, see : https://www.transition-bibliographique.fr/enjeux/bibliographic-transition-in-france/

[5] https://www.bnf.fr/fr/politique-de-catalogage-dans-bnf-catalogue-general

[6] https://www.bnf.fr/fr/politique-dindexation

[7] https://www.bnf.fr/fr/politique-de-qualite-des-donnees

JLIS.it

We should distinguish "registration" from "attribution". "Attribution" means that library attributes directly a given identifier, following international policies and rules. This is the case for ISSN, through ISSN French Centre, which *attributes* ISSN identifiers following rules and policies validated by the ISSN International Centre and ISSN international network. "Registration" means sending data for asking attribution to international authoritative body. For instance, BnF *registers* ISNI by sending authority records for names of persons and the bibliographic records linked to them to the ISNI International Attribution Agency, by getting back ISNIs attributed on its own data by this attribution agency, and by disseminating ISNIs through the book supply chain and the library community in France.

The third tension regards the necessity to keep a two-fold diachronic, dynamic approach. On the one hand, the international landscape of identifiers is moving. On the second one, the responsibility libraries can take on one given identifier can move, too. For instance, BnF is thinking about taking more responsibility on ISAN, ISWC and ISRC, depending on their business model, legal structure, on the one hand, and on resources BnF can invest on them, on the other hand.

Therefore, setting up an identifiers' policy means to declare principles, on which BnF can commit itself, by taking into accounts these tensions, and concrete conditions allowing such a commitment by a State and non-for-profit institution to be concretely achieved.

The policy content

The key-principle is permanence. The identifier shall give guarantees on permanence, which concretely implies for it to be based on shared, transparent governance, broad and, if possible, global community, sustainable business model, as for the identifier itself, as for the community using it, and a standardizing process. All these elements create trust in the opportunity of consuming human and financial resources to integrate the identifier in the library dataflows and in the services and engagement the library agrees on with other players.

We have also formulated four main conditions to make these principles concretely applied.

1. The identifiers must benefit from a broad and stable community or inter-community commitment It implies that the identifier is part of a normative strategy:
   - either because it corresponds to an ISO standard (for example: ISO 27729: 2012 for the ISNI identifier; ISO 15706: 2002 and ISO 15706-2 for the ISAN identifier; ISO 3297 for the ISSN) and undergoes the international consultation process applied to periodically revised ISO standards;
   - or because it is part of a strategic standardizing process (for example: ARK[8])

   The identifier must therefore benefit from support of an international community or of a cross-domain commitment, depending on its scope of use. Its use must also be recognized and promoted by one or more communities. The governance of the identifier, whether at a national or international level, must be based on a written contract and allow the community or communities to be represented in decision-making bodies and to contribute to the technical and strategic orientations of the identifier.

---

[8] See above

JLIS.it

The identifier must be based on a negotiated, transparent, stable, contractual and sustainable economic model for a public institution. Business model should enable the BnF to develop a medium and long-term policy of use and services for the communities it serves. It must also be balanced in order to provide guarantees of financial stability in the medium term. This is the case, for example, for the ISNI business model, which allows libraries overall business model of this identifier.

2. The identifier must have a clear and explicit application policy, in other terms, we must clearly know what does identify the identifier. For instance, we know to what entity ISNI is applied for, as described in the ISO corresponding standard, which put forward the concept of "public identity", more or less similar to the concept of "bibliographic identity" libraries are familiar with.

   The identifier must respect the principle of uniqueness. An identifier relates to one and only one resource. When a resource is stable, so is the identifier. When a resource changes to become something else, a new identifier must be assigned. Similarity and duplication issues need to be identified and addressed. For ARK, BnF is developing practices of redirection when merging two duplicates, for instance. The question is more sensible for concepts and remains under discussion for now, because two concepts are never exactly similar.

   The identifier data model must be defined, documented and transparent. The attribution policy and the scope of data and resources to which the identifier applies must be stable, unambiguous and explicit. The conditions for attributing the identifier must be clear and explicit so as to control the mechanism and scope of their attribution, as well as their non-reassignment. This is why BnF has made explicit the scope of ARK, and has recently extend it to records for archives and manuscripts, so as to make every BnF resource covered by this identifier, without any regard to the data base describing it.

3. The identifier must be technically sustainable. The ID is built to last.
   That means:
   - It must be independent from the technical protocols to ensure its attribution and management, as well as of the authority that technically ensures its attribution. The guarantees of technical sustainability must be made explicit in the contractual commitments binding the national or international governance body on the one hand and the BnF on the other. This is the case for ISNI, for instance.
   - The link between the identifier and the resource described must be permanent. The existence of the identified resource must be certified. We are developing the scope of the future French National Entity file (FNE), to be published around 2024, in this direction. The entity, and the identifiers associated to this must correspond to a real resource belonging to a member of the FNE network.
   - An identifier must be maintained during and beyond the life of the resource that it identifies. If the resource or entity evolves, the persistent identifier must ensure a redirection to the most recent version of the resource or of the description of the entity to which it returns. The user must be informed of any significant change in the identified resource: deletions, replacements, merges, substantial modifications of the scope of the resource. The memory of the assignment of the identifier must thus be preserved.
   - An identifier is never and under no circumstances reassigned.

- In accordance with W3C best practices, it is better for an identifier to be expressed as an URI, as allowed by ARK, for instance.
4. The identifier must be open and neutral politically and technically.
   This means :
   - The identifier must be administered by an independent body contributing to the neutrality and uniqueness of the Web. It does not depend on exclusive mercantile interests that unilaterally could impose objectives, governance and an economic model incompatible with the requirements of a public institution. Dedicated and trained teams follow the attribution and registration procedures. This is the case with the ISNI governance structure and Quality Team.
   - The BnF favors identifiers that are opaque in their meaning in order to avoid the temptation to modify them if the resource or entity they identify changes and to allow their widest distribution.

Conclusion: audience and next steps

The identifiers' policy is intended to have both an internal and external audience. It aims at explaining cataloguers' and librarians the main directions BnF is implementing, and at committing BnF in its coming discussions with end-users and management bodies of identifiers. The next steps are to concretely develop this policy for the identifiers already used in the workflow.

An identifiers' policy shows how important identifiers are for the future of bibliographic control, by accelerating and making consistent the overall shift of data structure towards entity management. We could say it is both a tool for managing this shift and the aim this shift is supposed to achieve, because it is a tool to redraw the library role and place in the global data ecosystem. It supposes not to have a defensive approach but to elaborate strategic orientations for making libraries not a customer or a victim, but a genuine player in this shift.

JLIS.it

The International Standard Name Identifier: extending identity management across the global metadata supply chain

Andrew MacEwan[a]

a) The British Library

Contact: Andrew MacEwan, andrew.macewan@bl.uk
Received: 12 April 2021; **Accepted:** 3 June 2021; **First Published:** 15 January 2022

ABSTRACT

This article describes how ISNI is being adopted as a common identifier across disparate sectors of publishing. Whether publishing and distributing recorded music, film or text ISNI is making good identity management a staple element in the global metadata supply chain. As the content creation industries become more engaged with the value of embedding good metadata from the point of publication libraries can look forward to benefitting from a truly global revolution in the metadata supply flow. A case study describes how a British Library project has taken ISNIs already in the British National Bibliography and cross-matched them with data from UK publishers' own databases to embed ISNIs into the book supply chain. It also describes plans for ongoing publisher engagement through implementation of ISNI assignment into its cataloguing-in-publication workflows for UK legal deposit.

KEYWORDS

Authority control; Identity management; Identifiers; Names.

JLIS.it

Introduction

According to the standard ISO 27729 the International Standard Name Identifier was originally conceived as a "bridge identifier" with the ambition that it would be used for the identification of public identities of parties involved throughout the media content industries in the creation, production, management, and content distribution chains. This paper provides a brief update on how this ambition is beginning to be realised through the growth in adoption of ISNI in different publishing supply chains. Whilst this is important for the growing utility of ISNI in breaking down metadata silos in relation to efficient name identification it is also important to contextualise this as part of a broader trend that is seeing the business of producing well-controlled metadata become part of the business of publishing in the age of digital supply and demand. This paper, however, will focus on ISNI as an exemplar of this trend and will report in particular on a British Library case study describing our engagement with a group of UK book publishers and other agencies to embed ISNIs in the book supply chain.

Metadata silos and the supply chains

Different forms of creative content are distributed in supply chain metadata silos specific to each content type. The standards followed in each supply chain are well documented on websites promoting their use. Text publishing is supported by metadata supplied in the ONIX schema, with enhanced subject access through THEMA subject codes and additional product control provided in the form of trade identifiers: ISBN, ISSN, EAN barcodes, DOI, etc., as described at the EDItEUR website (EDItEUR, n.d). The music industry mirrors this with the DDEX schema standard, underpinned by the use of identifiers to express products at varying levels of granularity: ISWC, ISRC, RIN, RDR, etc. all described at the DDEX website. Metadata standards for the film industry are described most comprehensively at the website for the Entertainment Industry Identifier Registry (EIDR, n.d). Library standards have the advantage of attempting to accommodate and describe different content types in common standards, but even so libraries too have also worked in their disconnected silos reflecting historical divisions in curation of different content types. At the British Library our Sound Archive, our general catalogue, and our manuscripts and archives are catalogued in separate databases that reflect the major differences in the types of content and the standards that we use to describe them.

Library metadata itself exists in a silo in the context of the global supply chains. We rely on crosswalks and mappings, such as ONIX to MARC, to re-use data from the supply chain in our library based schemas. We also rely heavily on industry standard identifiers like the ISBN and the ISSN to build efficient automated workflows that allow machine matching based data enhancements from multiple sources. Co-operative cataloguing, the efficient re-use and sharing of metadata between libraries, where possible via automated workflows, is a staple activity fundamental to the efficient realisation of bibliographic control in the library world. In recent years the same theme of better metadata standards to support efficiency, automation and re-use have become a hot topic in every commercial supply chain in the publishing world. There is interest both in improving end-to-end metadata supply chains within each content industry and in building crosswalks between supply chains where appropriate commonality exists. In a Whitepaper on identifiers for artists

JLIS.it

(Movielabs, 2019), the company Movielabs reviewed existing approaches to name identification such as VIAF, ORCID and ISNI as potential models for managing identities for the film industry. The paper notes that the widespread adoption of ISNI in the music industry is a factor recommending ISNI adoption in the film industry, given high levels of commonality linking the sectors, rather than pursuing invention of another name identification standard.

An early example of building better metadata solutions around commonality was the collaboration on the "RDA/ONIX Framework for Resource Categorisation" (JSC-AACR, 2006) that connected the work of the revision of the Anglo-American Cataloguing Rules with the development of the ONIX standard in the publishing industry. In 2014 the Linked Content Coalition published a paper, "Principles of Identification" (Paskin & Rust, 2014) that highlighted the content neutral potential of ISNI as a name identifier that could be used across multiple supply chains. Most recently the UK standards body, Book Industry Communications, has launched a Metadata Capability Directory (Matthews, 2020) to promote and improve the use of metadata standards in the end-to-end text publishing supply chain. The Directory is intended to be a platform where the use of standards across the supply chain can be compared, deficiencies and opportunities identified, and collaboration on solutions initiated. In the music industry the by-line on the DDEX website perhaps best summarises the conversations and initiatives that are taking place in every supply chain: "DDEX is a standards setting organisation focused on the creation of digital value chain standards to make the exchange of data and information across the music industry more efficient." (DDEX, 2021)

This brief outline of the wider supply chain serves to highlight ISNIs place in the digital ecosystem of the global supply chains, but it also serves as a reminder that library metadata exists in the context of those supply chains and has the potential to benefit from the growing commercial interest in making metadata work better.

ISNI's place in the supply chain

The focus of the rest of this paper is on ISNI as a specific exemplar of a content neutral standard for name disambiguation that is starting to fulfil its purpose as a bridge identifier across sector specific silos for metadata. The foundation of ISNI in library metadata means that it already provides identification for authors, musicians, actors, editors, producers, artists and supports identification of both individuals and groups or organisations. In recent years, adoption has been strongest in the library sector and the music sector, with building blocks in place to encourage more widespread use in the book supply chain. ISNI's ability to work across so many specialist domains is based on a hub and spoke model in which Registration Agencies and Members provide sector expertise but work with a common database in the ISNI Assignment System, maintained by OCLC.

ISNI at work in the supply chain

In the music industry the ISNI membership list is growing. YouTube, Apple, Spotify and both major and minor record labels are set to be users of ISNIs and a growing network of music metadata

JLIS.it

organisations specializing in rights, credit and attribution of content to artists and performers are providing the engine rooms for the supply of ISNIs to the music industry. Currently listed on the ISNI website from the music sector (alongside YouTube, Apple and Spotify) are SoundExchange, Quansic, Qanawat, Consolidated Independent, Jaxsta, @Musiekweb, Muso.AI, The ISRC Team and Soundways. (ISNI, n.d.) The last of these, Soundways, is a sound engineering company that has built an ISNI Registration Service within its Sound Credit system. Soundways describe it themselves on the ISNI website: "Sound Credit's ISNI registration system is part of its larger system for music crediting, using Sound Credit's new massive cloud profile feature. Once music creators and engineers set up a free profile, they can be instantly credited simply by entering an email address, swiping a card at a kiosk, or selecting their profile in an app. Any credited profile in Sound Credit will automatically attribute their ISNI code to every project involving that creator, along with other identifier codes such as the IPI/CAE or IPN that users can optionally enter" (Sound Credit, 2020). The interface with the ISNI central database emphasizes search and entering rich metadata to ensure that each ISNI is unique in the central database whilst local control of identities is maintained in the Sound Credit system itself.

An example of similar intention in the book publishing industry came in January 2020, when the Frankfurt-based technology and information provider MVB took on the role of an ISNI RAG operating in Germany, Austria and Switzerland. The first step will be to assign automatically an ISNI to all creators listed in the Verzeichnis Lieferbarer Bücher (VLB), the books-in-print catalogue used in the German-speaking world. In a second step, publishers whose books are listed in the VLB will be able to register new ISNIs for the creators of their works – directly from the catalogue, and free of charge. (MVB, 2020)

The British Library and ISNI

The British Library has a long standing involvement with ISNI from being a member of the ISO 27729 International Standard Name Identifier Committee to draft the standard to becoming one of the Founding Members of ISNI acting jointly with the Bibliothèque nationale de France to co-represent the Conference of European National Librarians (CENL) on the ISNI Board. Working with the Bibliothèque nationale and OCLC we supported the foundational work to build the initial ISNI database from VIAF and other data sources. The BL and the BnF have continued to provide quality assurance services to the ISNI International Agency for the ongoing maintenance of the ISNI database.

When the British Library became an ISNI Registration Agency in its own right this marked a strategic shift in our goals for authority control away from name disambiguation in the British National Bibliography (BNB) and in our catalogues towards bridging data silos and exploiting the potential of a numeric identifier to build and embed identity management into the supply chain. There are three guiding principles for our implementation of ISNI:

1. Embed ISNI in all our cataloguing workflows
2. Automate processes as far as possible
3. Engage with the supply chain

Pursuing these principles involves overcoming significant challenges. The British Library's cata-

JLIS.it

loguing workflows with regard to authority control use the LC/NACO file. We hold a complete mirror copy of the LC/NACO file in our Aleph cataloguing system and maintain currency with the other LC/NACO nodes through daily file exchanges. Integrating ISNI into our authority control workflows will require ISNIs to be uploaded into this LC/NACO shared resource. Conversations and planning for this to happen at scale are ongoing with the Library of Congress and the Program for Cooperative Cataloging, but it is evident that capturing and loading all the ISNIs already associated with NACO records within the ISNI database will take place over an extensive time period. In the meantime we have focused on getting ISNIs into our legacy bibliographic data and engaging with the UK publishing supply chain. Happily these two endeavours have worked in concert as will be described below.

A British Library case study in supply chain engagement.

Serious engagement with publishers and other actors in the UK supply chain was initiated in two facilitated meetings in early 2018. In January 2018 Publisher Licensing Services, an organization providing collective licensing and rights management services for the publishing sector, and an ISNI member organization, hosted a meeting for publishers to discuss the potential use of ISNI for improving identification of publishers and imprints in the supply chain. This discussion led to a follow up meeting in March hosted by Book Industry Communication to explore the wider topic of ISNI for authors, publishers and imprints. A colleague from the Bibliothèque nationale joined this meeting to give a presentation on their integration of ISNI into their cataloguing-in-publication workflows for French legal deposit. Thanks to further advocacy and promotion by EDItEUR the interest sparked by both these meetings led to the establishment of an informal UK Publishers Interest Group comprising the following organisations:

> Bibliographic Data Services (BL's CIP subcontractor)
> Book Industry Communication
> British Library
> Cambridge University Press
> EDItEUR
> Hachette UK
> International ISBN Agency
> Harper Collins
> ISNI International Agency
> Nielsen Book (UK ISBN Agency)
> Pan Macmillan
> Penguin/Random House
> Publisher Licensing Services
> Bloomsbury

Early on the group settled on a remit to explore practical solutions for disseminating ISNIs that were already established in the ISNI database into bibliographic product records that were already held in common by publishers and aggregators and the British National Bibliography. It was agreed that the quickest way to demonstrate value at scale and to introduce ISNIs into the supply

JLIS.it

chain was to exploit what ISNI had already achieved in building its database of identifiers. Since the group as a whole had many different levels of capability for handling varieties of ONIX and MARC data it was also settled upon to make CSV files the medium of exchanging data between the British Library and the publishers themselves.

The starting point for the work was to get ISNIs into the British National Bibliography. Names in records in the BNB are the established name forms found in the LC/NACO file. We already had staff experienced in working with the Virtual International Authority File (VIAF) to associate VIAF and NACO IDs with the Linked Open Data version of the BNB. We also already had established links from ISBNs for product records, names in those records and LC/NACO IDs. By using the VIAF links we were able to pull across all ISNI-LC/NACO associations already established in VIAF clusters and bring the ISNIs back into the BNB. This provided us with a base file of 3,160,908 names in BNB records with assigned ISNIs for working with publishers' product data.

Each of the publishers in the working group provided us with sample files and later full back files as we developed the matching processes. Publishers provided us with a name string, their proprietary in house author ID, and its associated products. ISBNs were the key match point for identifying the target records and our staff developed algorithms to ensure we associated only confident matches between the LC/NACO name string and the publisher's name string to assign the corresponding ISNI. Differences between original publisher data and BNB catalogued data meant there were a variety of issues to work with: different name forms, punctuation and character set issues, reverse name forms, presence or absence of names for translators or illustrators, multiplicity of product ISBNs for the same work. The process was refined over time. Early results were quite variable between publishers and percentages of assignment relatively low in the first round of work. After several iterations and an expanded group of publishers' files to work with the latest results are as given in the table below.

Publisher	Number of names	Number of matches	Success rate
Atlantic	1,201	954	79%
Bloomsbury	43,558	28,420	65%
BurleighDodds	681	35	5%
ChannelView	1.392	1,146	82%
Canongate	521	363	70%
Cambridge University Press	21,298	16,292	76%
Dorling Kindersley	2,103	1,409	67%
Hachette	10,857	7,820	72%
Harper Collins	13,406	8,107	60%
Kogan Page	1,117	708	63%
Liverpool University Press	1,498	1,064	71%
PanMacmillan	1,642	1,332	81%
Penguin	14,297	9,638	67%

JLIS.it

Publisher	Number of names	Number of matches	Success rate
Pluto	1,589	1,107	70%
Random House	24,060	16,127	67%
Taylor&Francis	107,871	68,878	64%
Total	**247,091**	**163,400**	**66%**

Fig. 1. Publishers' Author Name Matching Results

Generally, we have achieved a high level of consistency in the results and feedback from those publishers who have integrated the ISNIs into their own databases has confirmed the accuracy of the assignments from their side. An additional benefit that has come out of the work is cross deduplication of authors between publishers and in some instances deduplication within a publisher's own author file. The figures for deduplication are as given in the table below (Figure 2).

Publisher	Number of de-duplicated IDs (across all publishers)	Number of de-duplicated IDs (within publisher)
Atlantic	0	12
Bloomsbury	5409	619
BurleighDodds	0	0
ChannelView	0	0
Canongate	139	0
Cambridge University Press	3363	2225
Dorling Kindersley	398	62
Hachette	1940	746
Harper Collins	2350	119
Kogan Page	0	0
Liverpool University Press	0	4
PanMacmillan	524	33
Penguin	3449	846
Pluto	0	4
Random House	4386	1788
Taylor&Francis	8650	9916
Total	30608	16374

Fig. 2. Publishers' Authors Names Deduplication Results

JLIS.it

Whilst the deduplication across publishers was an anticipated benefit of sharing a common supply chain author identifier, the cleanup of duplicates within a publisher's own data was an unexpected bonus, but one that demonstrated additional value in working across data silos. A further early bonus of this project with publisher data is the first example of a provided ISNI being re-used by Harper Collins in an ONIX record for a new publication by one of their authors. (Figure 3)

```xml
<TitleElement>
    <TitleElementLevel>01</TitleElementLevel>
    <NoPrefix/>
    <TitleWithoutPrefix textcase="02">Boy Giant</TitleWithoutPrefix>
    <Subtitle>Son of Gulliver</Subtitle>
</TitleElement>
<TitleStatement>Boy Giant: Son of Gulliver</TitleStatement>

<Contributor>                                      <Contributor>
    <SequenceNumber>1</SequenceNumber>                <SequenceNumber>2</SequenceNumber>
    <ContributorRole>A01</ContributorRole>            <ContributorRole>A12</ContributorRole>
    <NameIdentifier>                                  <NameIdentifier>
        <NameIDType>01</NameIDType>                       <NameIDType>01</NameIDType>
        <IDTypeName>HCP UK Author ID</IDTypeName>         <IDTypeName>HCP UK Author ID</IDTypeName>
        <IDValue>4121</IDValue>                           <IDValue>1897</IDValue>
    </NameIdentifier>                                 </NameIdentifier>
    <NameIdentifier>                                  <NameIdentifier>
        <NameIDType>16</NameIDType>                       <NameIDType>16</NameIDType>
        <IDValue>0000000121251907</IDValue>              <IDValue>000000012147035X</IDValue>
    </NameIdentifier>                                 </NameIdentifier>
    <PersonName>Michael Morpurgo</PersonName>         <PersonName>Michael Foreman</PersonName>
    <PersonNameInverted>Morpurgo, Michael</PersonNameInverted>   <PersonNameInverted>Foreman, Micha
    <NamesBeforeKey>Michael</NamesBeforeKey>          el</PersonNameInverted>
    <KeyNames>Morpurgo</KeyNames>                     <NamesBeforeKey>Michael</NamesBeforeKey>
</Contributor>                                        <KeyNames>Foreman</KeyNames>
                                                   </Contributor>
```

Fig. 3. Example ONIX record containing ISNIs

Future work with publishers

The above results reflect the work we have achieved so far but the UK Publishers Interest Group continues to meet and we have more work to do. Although we do not think we can achieve much more improvement in the match rates through further improvements to our matching processes there may be improvements to be gained via more direct work with the ISNI database itself. Although the ISNI database began its life with a series of regular uploads of relevant records from the full VIAF database the last of these took place in 2016. Since then ISNI has worked with direct authority file loads from the increasing numbers of national libraries who have joined the ranks of the ISNI membership. The British Library has recently completed work on preparing an update file from its own copy of the LC/NACO database from 2016 to the present for submission to the ISNI database to bring LC/NACO up to date in the ISNI assignment system. Where possible this was enriched by associating title and ISBN data with the LC/NACO records extracted from the BL's own catalogues and the LC Books All file to facilitate the matching and the rich record assignment processes in the ISNI Assignment System. Following this load the total number of assigned ISNIs associated with a LC/NACO identity stands at 5,553,823 persons and 602,288

JLIS.it

organisations. The results from this update will be used to re-run and fill some of the gaps in the publishers' results.

Following the above step the dialogue with the publisher group will move onto another stage. The high assignment rates already achieved have already opened up a conversation around and an appetite for 100% ISNI coverage in publishers' data. There are several avenues to explore for achieving this. The work to date has been an experimental project with the goal of seeding ISNIs at scale into the databases at the beginning of the supply chain. It has also been a mutually beneficial project to both publishers and the British National Bibliography. If we have run out of automated means to populate both the BNB and the publishers' databases then one option to provide a more intensive, manual level of intervention to fill the gaps could be a priced service for the remaining legacy data, acting in our role as an ISNI Registration Agency.

The other unresolved question though is the provision of new ISNIs for future authors. We are working on two solutions for this. One is the provision of a Registration Service portal for individual ISNI assignment requests. The second is the integration of ISNI assignment into our CIP workflows for our Legal Deposit intake and the development of a feedback loop to publishers along the lines already pioneered by the Bibliothèque nationale. Since the integration option accords with our lead principle of embedding ISNI into our workflows the steps to that are described first before concluding with an outline of the functionality of the Registration Service Portal.

Building a CIP workflow for ISNI assignment – next steps

As already noted earlier the British Library's cataloguing-in-publication programme has been contracted out and for many years has been provided by a company called Bibliographic Data Services (BDS). BDS have been a member of the ISNI UK Publishers Interest Group since its inception and they are closely engaged with the goal of embedding ISNI as a supply chain identifier. As part of the current CIP workflow BDS use and supply name headings from the LC/NACO file in their records. As a precursor to implementation of ISNI the British Library has already supplied a reconciliation file for corresponding LC/NACO – ISNI equivalents for BDS to use to facilitate automatic assignment on the back of their use of LC/NACO. The next step will be to update this file with an additional correspondence file based on the forthcoming update of the LC/NACO file in the ISNI database. Once this is in place BDS systems and workflows are primed and ready for implementation. At this point in time the details of a feedback loop to the publishers supplying BDS with pre-publication information to inform CIP work has yet to be determined, as does the potential role for BDS acting as a Registration Agency for original assignments, but the workflows as building blocks to inform those decisions will be in place.

Providing an ISNI Registration Self-Service Portal

The final piece of the British Library's engagement with the supply chain has been the development of an online service for individual requests. As part of our provision of quality assurance services to the ISNI-IA we respond to user queries and feedback, often leading to requests for updates and additions to existing records and requests for new assignments. We have firsthand experience of a wide

193

JLIS.it

level of interest in ISNI amongst smaller publishers and directly from authors, artists, and performers across all repertoires of creative content. We are also acutely aware that at this level of interest and engagement the fact that only ISNI Registration Agencies and ISNI Members can register new ISNIs through privileged access to the ISNI assignment system interfaces is a barrier for those without the means to engage at the membership level. Since we are dealing with requests of all kinds at an individual level that is growing alongside the broadening engagement in ISNI we also know it is cumbersome and costly to deal with these individual requests sent to us through system-generated emails.

As part of another project, involving cross-sector engagement with the music industry initiated by the British Library's National Sound Archive, the Mellon Foundation provided us with specific funding to build an End User Portal for ISNI assignment requests. We have now completed development of thus system and it became operational in February 2020. The portal supports three main functions: Search, Request, and Add Data. The portal mediates these functions to interact directly with the ISNI system. The first two of these mirror the capability developed recently by Soundways in their Sound Credit system described above. As with the Sound Credit system the BL Service requires users to register on the system to access the functionality. Search is the critical first step to ensure that a pre-existing ISNI is not over looked before submitting a request for a new ISNI. As a second check, when a request is submitted, it passes through the matching algorithms in the ISNI system in case a similar name identity does already exist. The Add Data function is an additional aspect of the service that will allow the many end users who want to enrich an ISNI record with additional titles or links to do so directly and easily. Editing existing data is not permitted because the ISNI database is built from the metadata of its Members and contributors and only members can edit their own data in an ISNI record. The British Library will regularly monitor and quality assure all activity and transactions that go through the portal.

Concluding reflections

This paper has sought to present a short update on the growing adoption of ISNI as a name identifier supporting different metadata supply chains. Although only a selection from all the activities going on across the ISNI network of 65+ Agencies and Members, it has provided examples that highlight drivers behind the interest from the supply chains. Drivers that position engagement with ISNI in the broader context of a more developed interest in the value of high quality metadata as an essential component in supply chain management for commerce, discovery and the attribution of rights. It has sought to contextualize the implications of this for libraries and our common interest in bibliographic control by show casing just one approach, developed by the British Library, at engaging with the UK book publishing supply chain. We depend much on the supply of publishers' metadata but we have only had limited influence on bringing it into convergence with libraries' metadata requirements. Although authority control is only a single component of library metadata it has long been one of our most expensive metadata creation activities. Shifting the task of authority control into simultaneous management of ISNIs in the supply chain provides us with an opportunity to share that cost and to share its value. Metadata conceived and developed in the library sector, redefined as identity management, becomes a shared, common goal and the global supply chain becomes part of the solution.

JLIS.it

References

DDEX, n.d. Accessed June 2021. https://ddex.net/

EDItEUR, n.d. Accessed June 2021. https://www.editeur.org/

EIDR, n.d. Accessed June 2021. https://www.eidr.org/standards-and-interoperability/

ISNI, n.d. Accessed May 2021. https://isni.org/

Joint Steering Committee for Review of AACR, "RDA/ONIX Framework for Resource Categorisation". Last modified august 3, 2006. https://www.loc.gov/marc/marbi/2007/5chair10.pdf

Matthews, Peter. 2020, "Introducing the BICMetadata Capability Directory". EDItEUR website last accessed June 2021. https://www.editeur.org/3/Events/Event-Details/561

MVB. 2020. "MVB becomes an ISNI Registration Agency". Press release posted on ISNI website, January 2, 2020. https://isni.org/page/article-detail/mvb-becomes-an-isni-registration-agency/

Movielabs. 2019. "Creating a Talent Identifier for the Entertainment Industry". Last modified August 28, 2019. https://movielabs.com/talentid/Talent_ID.pdf

Paskin, N & Rust, G. April 2014. "Linked Content Coalition Principles of Identification". Linked Content Coalition website. http://doi.org/10.1000/287

Sound Credit. 2020. "Music industry ISNI registrations now free and automated". Press release posted on ISNI website October 23, 2020. https://isni.org/page/article-detail/music-industry-isni-registrations-now-free-and-automated/

JLIS.it

VIAF and the linked data ecosystem

Nathan Putnam[a]

a) OCLC, http://orcid.org/0000-0002-3984-3035

ABSTRACT

This article reviews the founding, current state, and potential future of VIAF®, the Virtual International Authority File. VIAF consists of an aggregation of bibliographic and authority data from over 50 national agencies and infrastructures, systems that follow different cataloging practices and contain hundreds of languages. After a short history of the project, the results of surveys for implementers of linked data projects on the use of VIAF data and provides suggestions for future use and sustainability.

KEYWORDS

RDF; Library Linked Data; VIAF; History.

JLIS.it

The Virtual International Authority File, known as VIAF, provides cultural heritage institutions and users with access to a combined, single international authority file with data from national libraries and infrastructures worldwide. VIAF contributors supply authority data that is matched, linked, and clustered with existing VIAF entities. VIAF allows researchers to identify names, locations, works, and expressions while preserving the regional language, spelling, and script preferences. There are more than 50 VIAF contributors from over 30 countries. The VIAF Council governs VIAF, which includes representatives from the contributor organizations and provides guidance on the policies, practices, and operations of VIAF.

This article begins with a short history of VIAF, including some current statistical information. It then discusses VIAF within the larger linked data ecosystem through several surveys conducted by OCLC. It concludes with a discussion regarding OCLC's continued support for VIAF, its use and potential integration into OCLC's shared entity management infrastructure, and recommendations for further research and investigation.

VIAF history and current use

History

In April 1998, the United States Library of Congress (LC), the German National Library (Deutsche Nationalbibliothek, or DNB), and OCLC wanted to test linking to each other's authority records for personal names as a proof-of-concept project. In August 2003, the LC, the DNB, and OCLC formed the VIAF Consortium in a written agreement during the International Federation of Library Associations and Institutions (IFLA) conference in Berlin, Germany. In October 2007, the National Library of France (Bibliothèque nationale de France, or BnF) joined the consortium. After this, the four organizations, assuming the role of Principals, had joint responsibility for VIAF with the three libraries contributing authority and bibliographic content, while OCLC supported the software and infrastructure. Other organizations later joined the consortium as Contributors, providing source files and expertise to advance the state of VIAF. Due to the proof-of-concept success, the Principles and Contributors sought a suitable long-term organizational arrangement for VIAF. After considering several options, the Principals and Contributors agreed to transition VIAF to OCLC, which was completed in April 2012 (Murphy, 2012).

As of September 2020, VIAF receives data from 56 sources and includes 172 million bibliographic records, 87 million authority records, and 33 million cluster records. Records are clustered, i.e., grouped together, when they represent the same thing but with data from different sources. VIAF stores both the source record and the aggregated cluster record. Figure 1 provides detailed statistics on source information, authority records by type, and top language representation. The comparison year in the figure uses OCLC's fiscal year, which runs from July to June. Interestingly, the top three languages within VIAF are of the three Principal institutions, LC, DNB, and BnF. VIAF also contains a range of authority records from the sources, including 10.5 million corporate authorities, 10.9 million geographic authorities, approximately 60 million personal name authorities, and 5.7 million title authorities, totaling 87 million.

JLIS.it
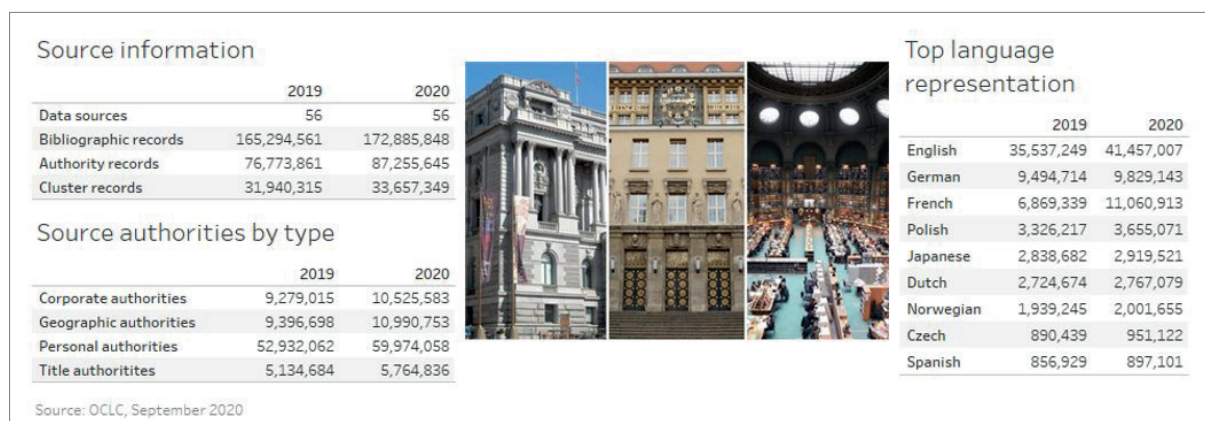
Fig. 1. Source, type, and language representation

VIAF continually adds data from existing and new sources. As seen in Figure 2, data clusters continue to grow, and the cluster types for personal, corporate, work, expression, and geographic authority records. While there are considerably more personal name clusters within VIAF, OCLC believes that the other types will increase in importance as existing and new users consume the VIAF data.
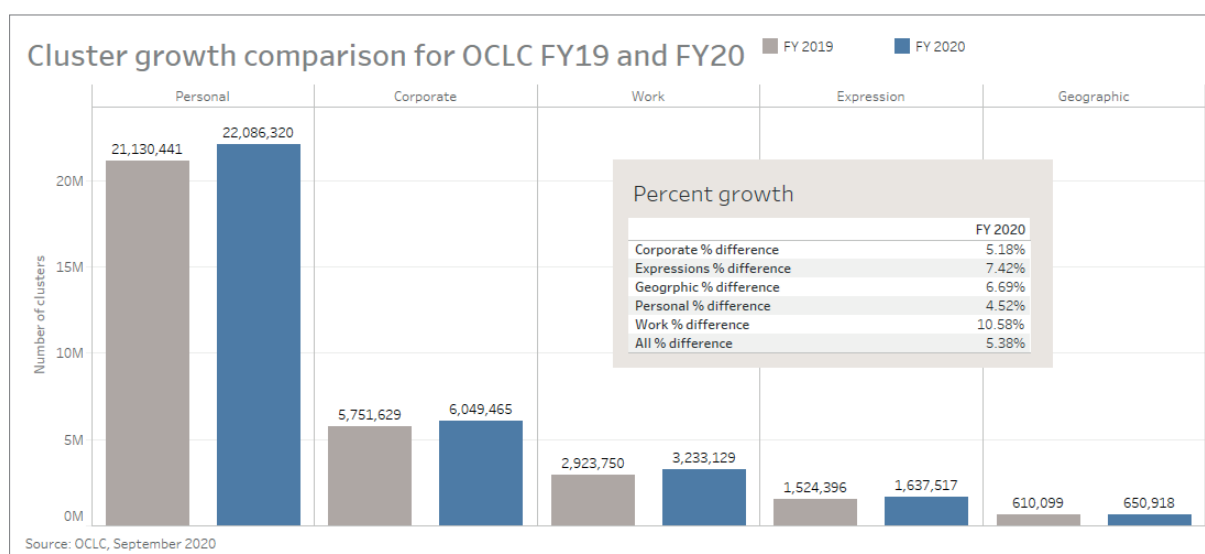
Fig. 2. Cluster comparison between OCLC Fiscal Year 2019 and Fiscal Year 2020

International linked data survey for implementers

During the past seven years, OCLC Research conducted surveys on implementing various linked data tasks and uses. Building upon the interest of the OCLC Research Library Partners Metadata Managers Focus Group, OCLC Research conducted the first "International Linked Data Survey for Implementers" between 7 July and 15 August 2014. This followed updates to the original survey in 2015 and 2018. This article discusses the results regarding the use of VIAF data, but analyses and

JLIS.it

results are available on the OCLC Research Linked data webpage (OCLC Research, n.d.). Interested persons can access the data directly on the OCLC Research linked data pages or through several articles written by Karen Smith-Yoshimura, including her discussion and analysis of the results[1]. Many institution types participated in the surveys including research libraries, national libraries, research institutions, library networks, governments, service providers[2], public libraries, museums, and a few classified as other. While research libraries continue to have many responders, Figure 3 shows the growing interest in different groups like national libraries, research institutions, and government institutions. Even on the lower end of the responder spectrum, public libraries and museums have seen a slight growth.
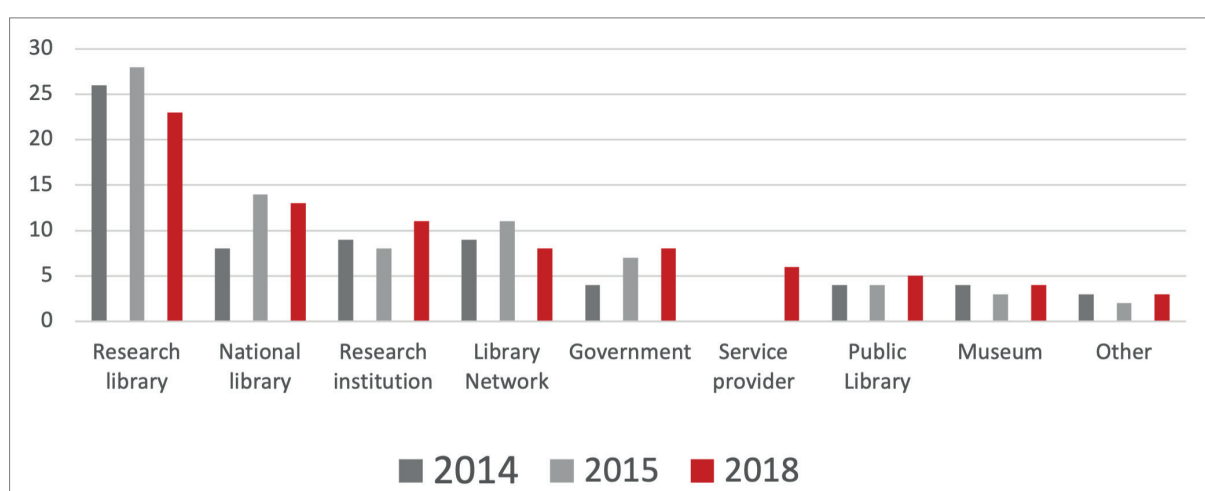
Fig. 3. Reponding insitutions by type

As the ecosystem matures and adoption increases, the participation of these groups will continue to grow. While there was a slight drop in the number of institutions answering how long they have had a linked data project or service in production, the number of projects and the time they have remained active continue to grow. 75% of the linked data projects/services described in 2018 are in production, slightly higher than the 67% reported in 2015. 40% of the linked data projects/services described in 2018 have been in production for more than four years.

The 2018 survey highlights the top seven linked data implementations. "Most used" is measured by the average number of requests per day, with all services reporting over 100,000 requests per day. All eight services have also been in production for more than four decades and include:

- American Numismatic Society's nomisma – a thesaurus of numismatic concepts[3]
- Bibliothèque nationale de France's data.bnf.fr – provides access to the BnF's collections and is a hub among different resources[4]

---

[1] See https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-survey.html for a complete listing of Karen's publications and presentations

[2] Service providers responded only to the 2018 survey

[3] http://nomisma.org/

[4] https://data.bnf.fr/

JLIS.it

- Europeana – an aggregation of metadata for digital objects from museums, archives, and audiovisual archives across Europe[5]
- Library of Congress Linked Data Service – provides access to over 50 vocabularies[6]
- National Diet Library's NDL Search – provides access to bibliographic data from Japanese libraries, archives, museums, and academic research institutions[7]
- North Rhine-Westphalian Library Service Center (hbz) Linked Open Data service – provides access to bibliographic resources, libraries and related organizations, and authority data[8]
- OCLC's Virtual International Authority File (VIAF) – an aggregation of over 50 authority files from different countries and regions[9]

Figure 4 shows the top ten linked data sources consumed by the 2018 survey respondents compared to 2015. The count of respondents in the 2018 and 2015 surveys was the same, 69 and 68, respectively. Six of the ten sources dropped between the 2015 and 2018 surveys while the other four grew. The most considerable change was in the increased use of Wikidata. And even though VIAF dropped between the two surveys, it is still ranked relatively high, coming in second after the Library of Congress's ID service.
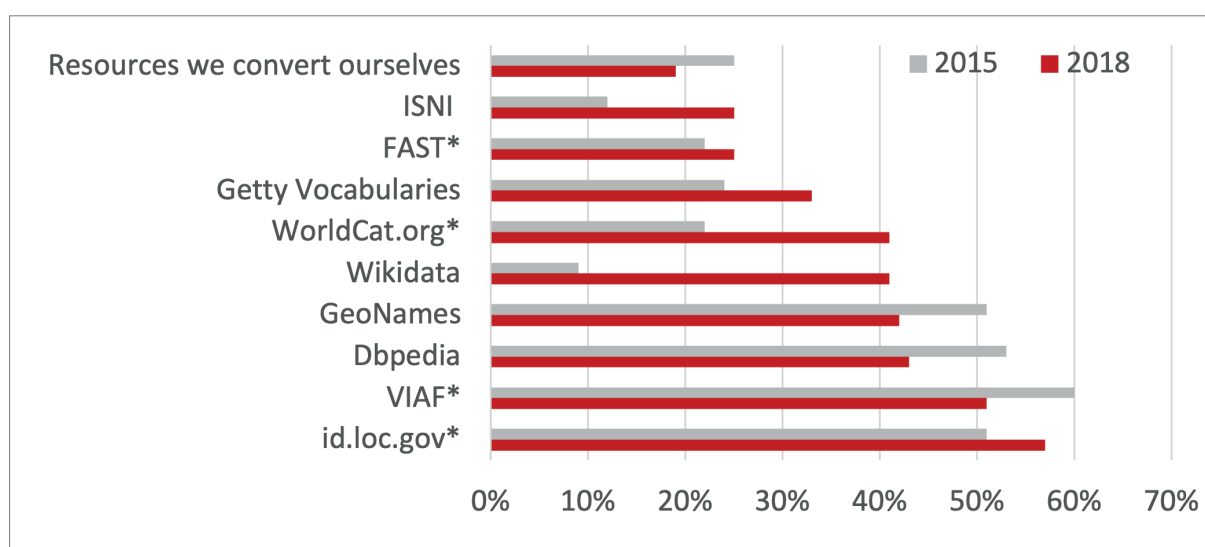
Fig. 4. A comparison of linked data sources consumed in 2015 and 2018

---

[5] https://www.europeana.eu/
[6] https://id.loc.gov/
[7] https://iss.ndl.go.jp/
[8] http://lobid.org/
[9] http://viaf.org/

JLIS.it

The potential for VIAF

VIAF continues to be supported by OCLC and, as discussed earlier, continues to add new sources and data. The ongoing success of VIAF for various consumers, including OCLC, will depend on greater integration into the linked data environment. Success includes transforming VIAF from a primarily MARC-based system to native RDF and integrating with RDF services and support. With financial support from the Andrew W. Mellon Foundation, OCLC is building a shared entity management infrastructure for library linked data. When completed in December 2021, this infrastructure will include authoritative descriptions of several types of entities, including works and persons, and will be enhanced and managed by the library and OCLC. Connections to other external vocabularies will place library collections in a broader context across the web.[10]

VIAF plays an integral role in the entity infrastructure, especially during the infrastructure's initial development phase. The grant-funded portion using VIAF entities to connect person entities within the infrastructure. During the first six months of data loading, selected VIAF clusters had connections to either WorldCat® works or Wikidata entities. The key to the initial phase was that the entities had built-in relationships with other entities that provided an enriched experience. The second six-month checkpoint continued the enrichment by adding additional entities. The second six-month checkpoint, which ended December 2020, included personal name entities from VIAF and work entities from WorldCat FRBR clusters. While not part of the grant requirements, it also had place entities from a separate linked data pilot for OCLC CONTENTdm®[11] with data from GeoNames, a database of geographical place names[12].

Areas for further investigation

The existing VIAF infrastructure continues to meet the goals of the original Principals and current Contributors. As with any library data project, continued usefulness will require change.

Two key areas to help determine the future of VIAF include running the implementers survey in the coming year and continued integration within the entity infrastructure. The implementer survey would indicate the continued use of VIAF within the larger linked data ecosystem and the probable and continued growth of Wikidata. Implementing VIAF into the infrastructure will help ensure stability and continuity as the ecosystem moves from record-based description to graph-based. Note that OCLC remains committed to providing a level of free access to those that wish to use the VIAF data regardless of in which ecosystem it finds itself. Additional areas for consideration include continued work with Wikidata partners to find solutions to challenges and the ever-on-going issues revolving around data quality and integrity.

---

[10] More information on the entity management infrastructure can be found at oc.lc/sharedentitymgmt.

[11] More information on the CONTENTdm Linked Data Pilot can be found at oc.lc/transform-linked-data.

[12] https://www.geonames.org/

JLIS.it

References

Murphy, Bob. 2012. Virtual International Authority File service transitions to OCLC; contributing institutions continue to shape direction through VIAF Council. 4 April. Accessed January 24, 2021. https://worldcat.org/arcviewer/7/OCC/2015/03/19/H1426803137790/viewer/file1365.html.

OCLC Research. n.d. Linked data from OCLC Research. Accessed January 24, 2021. https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-survey.html.

JLIS.it

Call me by your name:
towards an authority data control
shared between archives and libraries

Pierluigi Feliciati[a]

a) Università degli studi di Macerata, Dipartimento di Scienze della Formazione, dei Beni Culturali e del Turismo,
http://orcid.org/0000-0002-2499-8528

ABSTRACT

An important and not often addressed topic – considering the issues opened by cross-disciplinary projects – is the shared control of authority records, or better authority metadata, extended to other documentary and cultural heritage sciences. This paper will examine the potential opened by multi-dimensional and networked logics in the representation of entities in the form of data towards which the document communities are converging. This approach is even more valid if we consider the users' point of view, presently forced to jump from one information environment to another, and confront different names, forms and attributes for the same entities. The core entities to work on are persons, corporate bodies, places, chronological contexts, events, qualifying their relationships. After a brief resume of archival description's peculiarity, the paper highlights the updated standards available, mostly IFLA-LRM and RiC, precious documents to start from and stimulate an active collaboration. To facilitate the sharing, control, and enrichment of authority data in the form of RDF assertions, librarians and archivists may follow several pathways: matching the existing conceptual models, converging on a shared data playground like Wikidata, and developing foundational meta-ontology.

JLIS.it

Introduction: convergences between archives and libraries

In the digital era we live in, and after centuries of applying the profession in archives and libraries, documentary disciplines share some fundamental lines. For example, for preserving paper documents and records, quality and digital resources management, digital preservation, administrative metadata. However, there are traditionally few convergences about principles, methods, and informational approaches. The description seems to be the crucial activity that keeps the two professions furthest away, especially in Europe and mainly in Italy. Whether some bridges were more comfortable to be built, the informational approaches are commonly distinct because of the objects' nature, the separated communities and projects, and the awkwardness in converging towards shared goals. Nevertheless, this paper argues that it is impossible to postpone the goal of a shared, integrated control of authority data, extending the most up-to-date approaches to all the areas of documentary and cultural heritage disciplines. This paper focuses on the potentials of collaboration opened by the multi-dimensional and networked logic in representing information entities towards which the documentation communities are converging. Moving from the presentation of archival description peculiarity, matched with the recent evolutions for bibliographic catalogues, this paper will try to shape the future possibilities to activate the development and control of shared authority datasets.

Archival description and authority control

Traditionally, archival description produces closed information pieces, inventories, or finding aids, representing individual archival fonds, informing about their provenance and internal logical partitions (Duranti 1992). The descriptive standards released by the ICA-International Council of Archives from the 90s to 2008 formalized this approach at the international level. The sage, secular principle of *respect des fonds* provides that every fond has to be managed and described separately, as a particular case due to its creator's unique activity. Moreover, the multilevel description rules state that each level of description has to give «information for the parts being described», and archivists should «present the resulting descriptions in a hierarchical part-to-whole relationship proceeding from the broadest (fonds) to the more specific» (ICA 2000, 12). The context prevails over the content, and rarely inventories reach the item level, offering data about records. This model has necessarily held back any connection among descriptions, isolating every pair creator/fond as a unique informational resource. These standard-compliant descriptions are produced mostly adopting relational databases and made accessible through a textual search on descriptive fields. Consequently, the archival description has not easily followed the World Wide Web's evolutions, markedly in the new century, whether the archival information entities are shaped as definitive records with closed hierarchical relations and hardly could keep the form of graphs, neither hypertextual, nor semantic-based.

Regarding the authority records, the archival access points according to the ICA descriptive standards are referred just to archives' creators (corporate bodies, persons or families). They have to be «based upon the elements of description» and their informational value «is enhanced through authority control» (ICA 2000, 9). The ISAAR(CPF) rules guide archivists in editing authority records, even establishing relations between them, under some defined categories:

JLIS.it

hierarchical, temporal, associative, family (ICA 2003, 21-22). We had to underline that those standards' combined effect led to the loss of the access points included in the traditional archival finding aids: personal or corporate bodies' names, places, subjects (*notable things*). Indeed, some crucial elements like names, dates, events, and places were conceived just as attributes of the units of description.
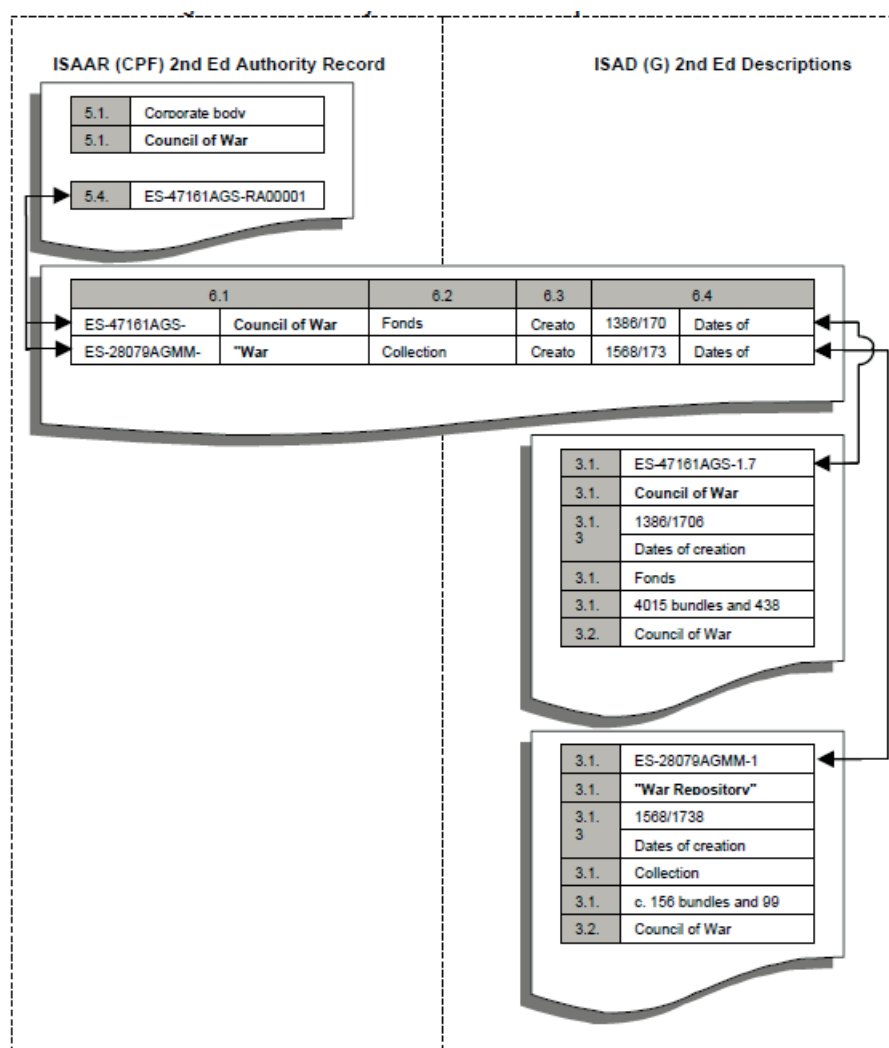
Fig. 1. Representation of how archival authority records can be linked with descriptions of archival materials (ICA 2003, 29)

Furthermore, names (i.e., units of description' titles) have to be extracted from archival files and other sources related to creators' internal organization, with the indication of limiting their normalization as much as possible. To explain better this traditional practice: suppose that the original name of an archival series is "Biccherne" (the magistrate or chancellery of finance from the 13th to the 14th century for Siena, Italy). Archivists are asked to describe this entity under "a formal title or a concise supplied title in accordance with the rules of multilevel description and national conventions." (ICA 2000, 14). This practice underlies two kinds of problems:

JLIS.it

1. interoperability and Autority Control: every description can only be checked by those who produce it, in possession of the bibliographical reference and especially of the wisdom arising from the heuristic study of the fond. It is almost impossible to build distributed authority control features, and the centralized control is allowed to verify the respect of formal rules;

2. users' friendiness: users may not know the fond or series's original name but are forced to query a database adopting Google-like behaviours. The adoption of relational databases caused the prevalence of searching vs. browsing services, and not-expert users may be deluded or lost while performing their research, as some studies clearly demonstrated (Duff, Stoyanova, 1998; Yakel, 2003; Chapman, 2010).

Nevertheless, recently some Linked Open Data data extraction experiments from archival DBs based on ICA standards were provided. Unfortunately, the assertions produced are not easy to be integrated into the Semantic Web *info-verse* because the ontologies adopted are local, representing specific data models, and could not be standardized in the absence of a shared Conceptual Model.

Archives in the *info-verse*: Records in Contexts

The new ICA standard RiC – *Records in Contexts,* defined by the EGAD – Experts Group on Archival Description from 2012 to 2016 turned upside-down the hierarchic and mono-dimensional logics of ISAD(G) and ISAAR(CPF). Proposing a multi-dimensional description, RiC Conceptual Model aims to be the reference for producing graphs of linked information entities instead of hierarchic or bare database rows connections. The 0.1 draft version of the Conceptual Model was published in August 2016 (ICA 2016) and questioned deeply by the international community (Bunn 2016; Duranti 2016; ANAI-ICAR 2017; SAA 2018). The recommendations covered several aspects: the "western" composition of EGAD, and the request to open RiC to existing ontologies like IFLA LRM (Riva et al. 2017), CIDOC-CRM (CIDOC CRM 2021), PREMIS (LoC 2018), and PROV-O (W3C 2013).

The draft version of RiC-CM was then updated in December 2019, publishing another draft version, the RiC-CM 0.2 (ICA 2019a), on which the RiC Ontology 0.1 (ICA 2019b), developed by the EGAD RiC-O team,[1] was based. Recently, in February 2021, the RiC-0 0.2 was released, compliant with the latest version of RiC-CM, 0.2, released in July 2021, and slightly different from RiC-CM 0.2 preview[2]. Again, a draft version explicitly to be corrected and enriched, in the perspective of the release of RiC-O 1.0. First of all, it "does not include the Conceptual Model Introduction, diagrams, or appendices". Moreover, it has to be quoted the absence of any explicit reference to the acceptance of the community's observations to the 2016 consultation draft and to the methodology adopted in the development process. As regards RiC-O 0.2, it lacks examples and tutorials, and it is explicitly declared that it "will continue to evolve, the next milestone being the release of RiC-O 1.0, which will probably take place by the end of 2021, at the same time as RiC-CM 1.0".[3]

---

[1] The EGAD RiC-O team is coordinated by Florence Clavaud (Archives nationales de France) and composed by Daniel Pitti (University of Virginia, USA), Aaron Rubinstein (University of Massachusetts Amherst, USA), Tobias Wildi (Docuteam GmbH, Switzerland) and Miia Herrala (National Archives of Finland).

[2] See https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf, accessed November 11, 2021.

[3] See https://www.ica.org/standards/RiC/RiC-O_v0-2.html, accessed November 11, 2021.

JLIS.it

Ric-CM 0.2 deeply changed the entities articulation present in version 0.1, adopting a four-level hierarchical logic: the macro-entity RiC-E01 *Thing* (first level) includes the entities RiC-E02 *Record Resource* (containing RiC-E03 *Record Set*, RiC-E04 *Record* e RiC-E05 *Record Part*), RiC-E06 *Instantiation*, RiC-E07 *Agent* (containing RiC-E08 *Person*, RiC-E09 *Group*, articulated in RiC-E10 *Family* and RiC-E11 *Corporate Body*, RiC-E12 *Position* e RiC-E13 *Mechanism*), RiC-E14 *Event*5 (specifiable with RiC-E15 *Activity*), RiC-E16 *Rule* (specificabile con RiC-E17 *Mandate*), RiC-E18 *Date* (specifiable with RiC-E19 *Single Date*, RiC-E20 *Date Range* or RiC-E21 *Date Set*), and RiC-E22 *Place* (see fig. 2). The entities and sub-entities of RiC-O are expressed as classes, and the properties are detailed in the datatypes. It has to be noted that the Internationalized Resource Identifier of RiC-O is not yet active, so it is not possible to refer to the namespace and allow applications to be automatically processed. This draft state of the new standard, and the Experts Group on Archival Description's isolation from the international community cannot help slow down the development of description tools based on RiC, any projects of conversion of existing catalogues, and the availability of archival linked triples in the semantic info-verse. Anyway, some isolated experiments, not ascribable directly to EGAD, started. We can quote the case presented in a spanish paper (Llanes-Padrón, Pastor-Sánchez and Juan-Antonio, 2017), the French proof of concept PIAAF, *Pilote d'interopérabilité pour les Autorités Archivistiques françaises* (Clavaud 2018),[4] and the Matterhorn RDF Data Model, based on RiC but open to existing ontologies (Dubois, Nef, 2017).

Another archival ontology to consider is the EAC-CPF Ontology (Mazzini, Ricci, 2011), based on the XML schema maintained by the Society of American Archivists with the Berlin State Library. It is used for encoding contextual information about persons, corporate bodies, and families related to archival materials, encoding the rules published in ISAAR(CPF). Some updated archival description applications are offering the export feature of RiC-like RDF triples, converting the hierarchical descriptive structures into multi-dimensional graphs. Nevertheless, nowadays, archives' global semantic interoperability is quite tricky without a wide-accepted, stable and accessible ontology.

Metadata integration between archives and libraries

The notion of catalogue could be taken in its broadest sense: ordered and systematic collection or record of items. Its function could not be reduced to the retrieval and identification of a single item, having the role of activating unexpected connections between different items:

> *Functions of the Catalogue*: The catalogue should be an efficient instrument for ascertaining 2.1 whether the library contains a particular book […] and 2.2 (a) which works by a particular author and (b) which editions of a particular work are in the library. (Statements 1961, 1).

Adopting this broad notion of catalogue, archival finding aids can also be considered catalogues (term commonly used in English). This phenomenon is even more reasonable considering that the outlines of informative objects tend to blur on the web, and in the web of data they are reduced to minimal assertions[5]. Considering the present tendencies in the archival and bibliographic de-

---

[4] See also http://piaaf.demo.logilab.fr/, accessed april 7, 2021.
[5] See Michetti 2020, 28, note 9.

scription, we may dare to say that both inventories and catalogues are conceptually and technically outdated. The documentary communities are asked to produce, control, share, monitor and enrich pieces of data, no more deep-web records, entrusting them to be accessed in the *infoverse*, understood, used and re-launched by human or web agents.

Authority control represents an important function to ensure the quality of linked open meta/data, produced through the intermediation of libraries networks but more e more in collaboration with the other memory institutions such as Archives and Museums. Firstly, it is no longer sustainable the management of authority control just at a local or national level. Then, the perspective must be broadened beyond the provenance descriptions, bibliographic, archival or relating to other human artifacts, such as artworks. While respecting the specificity of disciplines, the priority sandbox for archivists and librarians could be sharing authority data, giving to persons, agents, organizations, dates, places, and activities more knowledge facets. Despite the uncertainties, the road of data integration seems to be drawn. The approach driven by RDA and IFLA-LRM (Riva et al. 2017), jointly with the future, stable version of RiC-CM, could be the starting pillars to base on the collaboration. Several pathways to reach this goal could be followed. The first, maybe more manageable, is enabling the quoted conceptual models to talk, i.e. converging on the same concepts (entities) and defining the possible relations.

To open the work to be done, the Table 1 is a starting, tentative of matching the core entities of IFLA-LRM and RiC-CM 0.2. The RIC-E01 *Thing* is not that far from the *Res* entity of IFLA-LRM, considering their relations on the one hand with *Record Resource*, *Agent* , *Event*, and *Date*, on the other with *Work/Item* (considering the substantial unicity of records), *Time-span*, *Place* and *Agent* (Person, Collective Agent). The LRM conception of *Nomen* as an appellation of *Res* could be an interesting question to be addressed in the stable version of RiC, considering the complexity of appellations in archival description: original, derived, normalized, synthesized.
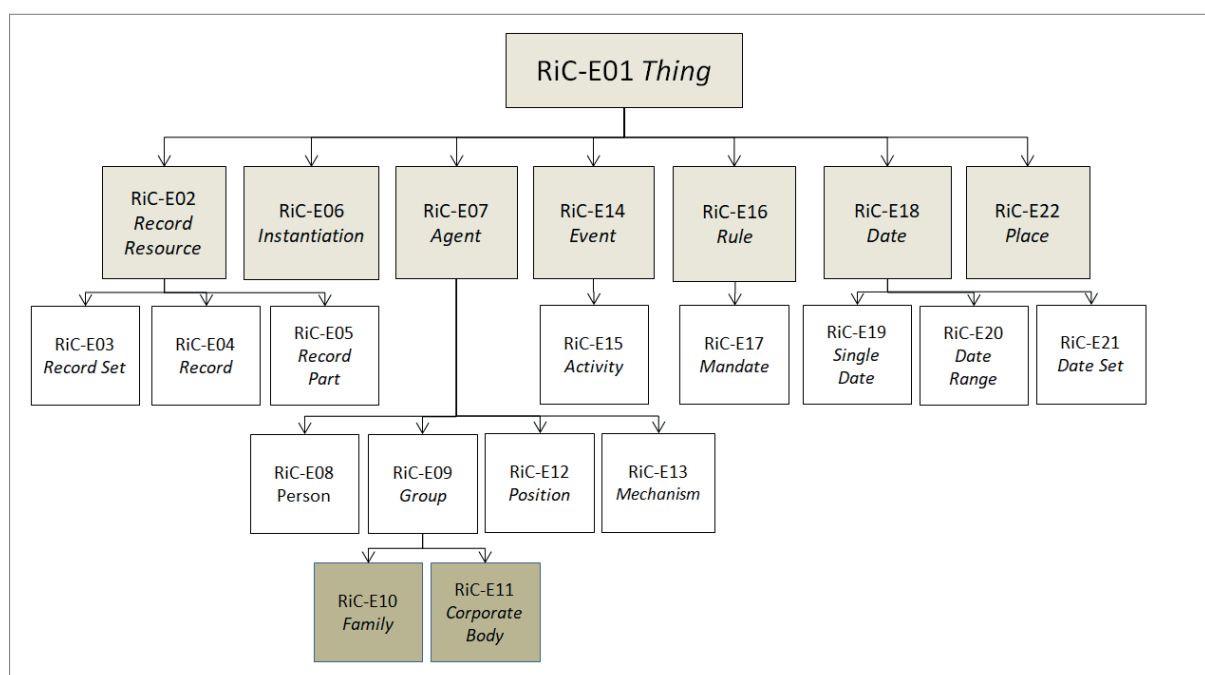
Fig. 2. RiC-O *diagram of entities* (Feliciati 2021, 99)

JLIS.it
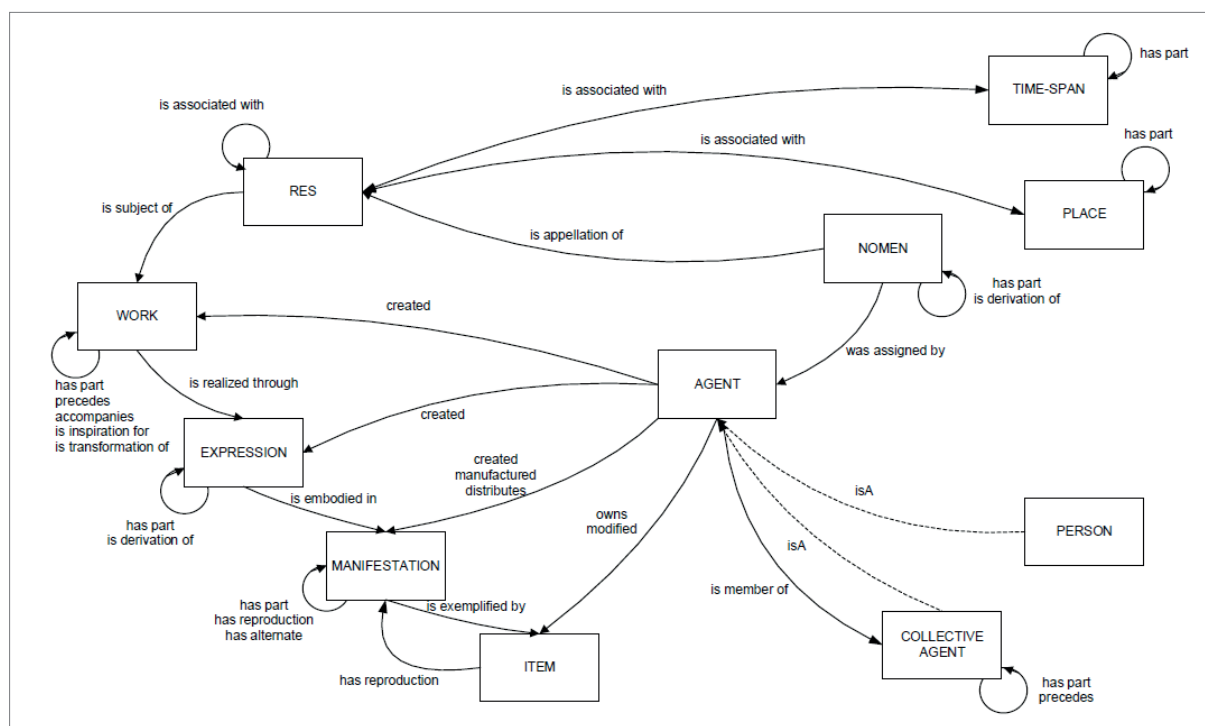
Fig. 3. IFLA-LRM table 5-6, *final overview diagram*

IFLA-LRM	RiC-O
Res + Nomen	Thing
Time-span	Date (Single Date, Date Range, Date Set)
Place	Place
Agent (Person, Collective Agent)	Agent (Person, Group, Position, Mechanism)
Work/Item	Record Resource

Table 1. Tentative correspondence between IFLA-LRM and RiC-O core entities

The second path to be followed is cooperating actively on a meta platform, a shared data playground, like Wikidata.

Wikidata (https://www.wikidata.org) is a project developed starting from its mother project, Wikipedia (https://www.wikipedia.org), both free and open repositories accessible over the web. Unlike Wikipedia, Wikidata stores information as structured data in a database. While the primary mission of Wikidata was to serve as a central repository for Wikipedia and other Wikimedia projects, it plays now the role of an independent, open, collaborative, and versatile platform. It could be used for «many different services and applications, from reusing identifiers to facilitate data integration, providing labels for multilingual maps and services, to intelligent agents answering queries and using background knowledge» (Vrandecic, 2013, p. 90). Wikidata uses Linked Open Data to store facts about items as nodes linked by properties as vertices; thus the project is often

JLIS.it

referred to as a linked open data repository of facts, available under an open CC 0 license. Tim Berners-Lee argued that his Semantic Web vision was hard to be realized because the ontologies must be developed, managed, and endorsed by (missing) practice communities. With Wikidata successfully serving as a LOD repository of facts, the Semantic Web's vision idea seems feasible. If we regard archives and library metadata as (functional) statements of facts that facilitate access of knowledge materials, Wikidata can be adopted as an ideal tool to make these facts accessible and discernable to machines and intelligent algorithms. In fact, «Wikidata can be used to make these facts accessible and discernable to machines and intelligent algorithms to realize the vision of the Semantic Web. For instance, it is quite conceivable to imagine that library patrons in the future may no longer use library catalogues and depend on intelligent devices and algorithms to search and access library holdings over the web» (Tharani, 2021, 2).

Many working groups of librarians are active on Wikidata managing and enrichment, defining a metadata structure for libraries and uploading and sharing local metadata globally (Bergamin, Bacchi 2018)[6]. Some archivists launched recently the *Wikidata:WikiProject Archival Description*, with the aim «to create the world's most comprehensive high quality database of archival fonds and heritage collections, to represent archival structures within Wikidata where this is deemed useful and to ensure the interlinking between archival finding aids and Wikidata»[7]. The project, connected with the *Wikidata:WikiProject Archives Linked Data Interest Group*, is leaded by French archivists and is considering the elaboration of ICA descriptive standards before RiC. In Italy, since 2020, is active the *Wikidata:Gruppo Wikidata per Musei, Archivi e Biblioteche* (GWMAB)[8], inspired by the Wikidata Affinity Group[9], launched mainly by librarians but open to the potentialities of Wikidata for Museums, Archives and Libraries. The purpose of this group to support culture professionals is going to produce some results in adding and correcting metadata related to museum and archives.

In order to figure out the shared work to be done on Wikidata, it could be useful the presentation of a case of possible trans-disciplinary integration: Umberto Eco. Umberto Eco (1932 –2016) was an Italian medievalist, philosopher, semiotician, cultural critic, political and social commentator, and novelist. After his death, his library is presently going to be split into two collections: the ancient books sold to Biblioteca Braidense (Milan) and his modern books and archival records, donated to the University of Bologna. The "Eco, Umberto" authority records in ISNI (0000 0001 2283 9390), VIAF (108299403), and other sources like the Italian SBN (CFIV006213) refer just to his being an author of works. Nevertheless, he was a library collector and owner, an archives creator, a subject of books and essays, of art portraits, photos. Besides the authority record and the Wikidata entity of interest concerning him, the places related to his life and work, the institutions holding his personal library and archives, his political activity, his family, his relationship with many other people should be semantically represented by letting different professionals working on the same information units. The Wikidata element referred to Umberto Eco (Q12807) [10], relatively poor at the time of the

---

[6] See https://www.wikidata.org/wiki/Wikidata:WikiProject_Libraries, accessed November 21, 2021.

[7] See https://www.wikidata.org/wiki/Wikidata:WikiProject_Archival_Description, accessed November 21, 2021.

[8] See https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche, accessed November 21, 2021.

[9] See https://wiki.lyrasis.org/display/LD4P2/LD4-Wikidata+Affinity+Group, accessed November 21, 2021.

[10] See https://www.wikidata.org/wiki/Q12807, accessed November 21, 2021.

JLIS.it

Bibliographic Control Conference, was enriched in the subsequent weeks. It attributes properties about his personal and professional life, his *notable work* (P800), *awards received* (P166), and his being the *owner of* (P1830) a personal library. The collaboration of archivists to enrich this element could add more properties, like his being a *creator* (Q59275219), *collection creator* (P6241), enrich the element *Umberto Eco's library* (Q35029860) and create the element referred to his archive.

The third pathway to build shared authority control between archivists and librarians could be the convergence towards a brand new foundational Conceptual Model.

The focus of this line of work could be the selection of shared classes, entities and properties, such as agents (persons, corporate bodies, families), their roles/functions in different contexts, geographic names (even historical), chronological data (exact dates or data range), actions/events, qualifying their multiple relationships. To develop this needful reference model and trans-ontology could facilitate and ebnable the integration of authority records in the form of RDF assertions. Collecting, connecting, enriching and controlling high-quality semantic information provided from different data sources will increase the potential of online services, making them richer and more useful for final users.

Conclusions

A shared approach to authority control would be even more valid considering the final users' perspectives. At present, as users, we are often forced to jump from one online source to another, even produced by the same institution, to compare and choose different forms of names and attributes referred to the same entities. Our time is not saved. The quality of use for documentary environments needs to be increased through an integrated approach to authority control and the adoption of updated metadata technologies. This strategy could represent a virtuous opening to the wisdom of crowds, by systematically sharing rich LODs, allowing users' annotations, using UX mining and collaborating with a global multilingual knowledge graph like Wikidata.

Interoperability should be possible with other cultural semantic sets of LODs, mostly produced by cultural heritage institutions different from archives and libraries. The goal could be the extension and enrichment of contexts and relations, representing the actual complexity of human activities in times, without reducing the semantic richness of descriptrive data. This perspective marks a step ahead compared with web portals, harvesting simplified metadata sets from data providers' repositories and necessarily affected by the issues of overwhelming search results. In this sense, the CIDOC-CRM model paved the way for semantic models in the cultural heritage sector. Any interoperability perspective can not help but compare with its classes and properties. The challenge posed by the semantic web forces the culture professionals to take a step forward in representing human activities. We have to break down disciplinary walls, enlarge the concept of provenance (Lemieux 2016) and respect the complexity, heterogeneity, discontinuity and transversality of contexts.

Some issues could slow down this process: organizational, the availability of models for standardization, the disciplinary edges. Some organizations, better if international, should take the initiative to launch this ambitious project by calling on experts from different sectors, archivists, librarians and cultural heritage experts to action. We have just to be ready to answer.

JLIS.it

References

ANAI-ICAR. 2017. "Records in Contexts. A conceptual model for archival description (draft v0.1, September 2016). Il contributo italiano", *Quaderni del Mondo degli Archivi*, 2 (luglio 2017), http://www.ilmondodegliarchivi.org/images/Quaderni/MdA_Quaderni_n2.pdf, Accessed April 6, 2021.

Bergamin, Giovanni; Bacchi, Cristian. 2018. "New ways of creating and sharing bibliographic information: an experiment of using the Wikibase Data Model for UNIMARC data". *JLIS.it*, v. 9, n. 3, p. 35-74, sep. 2018. http://dx.doi.org/10.4403/jlis.it-12458. Accessed April 13, 2021.

Bunn, Jenny. 2016. *Results of the ARA SAT consultation on Records in Contexts*, https://www.archives.org.uk/about/community/groups/viewbulletin/59-results-of-the-ara-sat-consultation-on-records-in-contexts.html?groupid=21. Accessed April 6, 2021.

Chapman, J., C.. 2010. "Observing Users: an Empirical Analysis of User Interaction with Online Finding Aids". *Journal of Archival Organization*, 8, 4-30 (2010), https://doi.org/10.1080/15332748.2010.484361. Accessed April 11, 2021.

CIDOC CRM Special Interest Group. 2021. *Definition of the CIDOC Conceptual Reference Model*. Version 7.1, March 2021, http://www.cidoc-crm.org/sites/default/files/CIDOC%20CRM_v.7.1%20%5B8%20March%202021%5D.pdf. Accessed April 6, 2021.

Clavaud, Florence, 2018. *Semantizing and visualising archival metadata: the PIAAF French prototype online*. May 4, https://www.ica.org/en/semantizing-and-visualising-archival-metadata-the-piaaf-french-prototype-online. Accessed April 6, 2021.

Dubois, Alain, Nef, Andreas. 2017. *The Matterhorn RDF Data Model: Implemeting OAIS and RiC in the context of semantic technologies*. Presentation, http://www.alaarchivos.org/wp-content/uploads/2017/12/3.-Alain-Dubois-Andreas-Nef.pdf. Accessed April 7, 2021.

Duff, Wendy, Stoyanova, Penka.1998. "Transforming the Crazy Quilt: Archival Displays from user's point of view". *Archivaria*, 45, 44-79 (1998), https://archivaria.ca/index.php/archivaria/article/view/12224. Accessed April 13, 2021.

Duranti, Luciana. 1992. "Origin and Development of the Concept of Archival Description". *Archivaria* 35 (January), 47-54, https://archivaria.ca/index.php/archivaria/article/view/11884. Accessed April 6, 2021.

Duranti, Luciana (compiler). 2016. Comments on "Records in Context". InterPARES Trust, https://interparestrustblog.files.wordpress.com/2016/12/interparestrust_commentsonric_final2.pdf, Accessed April 6, 2021.

Feliciati, Pierluigi. 2021. "Archives in a Graph. The Records in Contexts Ontology within the framework of standards and practices of Archival Description". *JLIS.it*, Vol. 12, No. 1 (2021), http://dx.doi.org/10.4403/jlis.it-12675. Accessed April 13, 2021.

ICA (International Council on Archives) – EGAD (Experts Group on Archival Description). 2016, *Records in Contexts. A conceptual model for archival description. Consultation Draft v0.1*, September, https://www.ica.org/sites/default/files/RiC-CM-0.1.pdf. Accessed April 6, 2021.

JLIS.it

ICA – EGAD. 2019a, *Records in Contexts. A conceptual model for archival description. Consultation Draft v0.2* (preview), December, https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf. Accessed April 6, 2021.

ICA – EGAD. 2019b, *Records in Contexts Ontology (ICA RiC-O) version 0.2*, 2019-12-12, https://github.com/ICA-EGAD/RiC-O/blob/master/ontology/previous-versions/RiC-O_v0-1_release/RiC-O_v0-1.rdf. Accessed April 6, 2021.

ICA – EGAD. 2021a, *Records in Contexts Ontology (ICA RiC-O) version 0.2*, 2021-02-12, https://www.ica.org/standards/RiC/RiC-O_v0-2.html. Accessed April 6, 2021.

ICA – EGAD. 2021b, *Records in Contexts. A conceptual model for archival description. Consultation Draft v0.2*, July, https://www.ica.org/sites/default/files/ric-cm-02_july2021_0.pdf. Accessed August 6, 2021.

ICA - Committee on Descriptive Standards. 2000. *ISAD(G): General International Standard for Archival Description, Second Edition*. Ottawa, https://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_EN.pdf. Accessed April 6, 2021.

ICA - Committee on Descriptive Standards. 2003. *ISAAR (CPF): International Standard Archival Authority Record For Corporate Bodies, Persons and Families. Second Edition*, https://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd. Accessed April 8, 2021.

IFLA – International Federation of Library Associations and Institutions, Cataloguing Section and Meetings of Experts on an International Cataloguing Code. 2017. *Statement of International Cataloguing Principles (ICP).* https://www.ifla.org/files/assets/cataloguing/icp/icp_2016-en.pdf. Accessed April 10, 2021.

Lemieux, Victoria (ed.). 2016. *Building Trust in Information. Perspectives on the Frontiers of Provenance*. Springer International Publishing. Senza luoigo?

Llanes-Padrón Dunia, Pastor-Sánchez Juan-Antonio. 2017. "Records in contexts: the road of archives to semantic interoperability", *Program*, 51:4, 387-405, https://doi.org/10.1108/PROG-03-2017-0021. Accessed April 6, 2021.

Library of Congress - PREMIS Editorial Committee. 2018. *PREMIS 3 Ontology.* https://id.loc.gov/ontologies/premis-3-0-0.html. Accessed April 6, 2021.

Yakel, E.. 2003. "Impact of Internet-Based Discovery Tools on Use and Users of Archives". *Comma*, 191-200 (2003).

Mazzini, Silvia and Ricci, Francesca. 2011. "EAC-CPF Ontology and Linked Archival Data", *Proceedings of the 1st International Workshop on Semantic Digital Archives*, September 29, 72-81, http://ceur-ws.org/Vol-801/paper6.pdf. Accessed April 6, 2021.

Michetti, Giovanni. 2020. "'Il mondo come puzzle': i beni culturali nel web". *Digitalia*, Anno XV, Numero 1 - Giugno 2020, 26-42, http://digitalia.sbn.it/article/view/2485. Accessed April 6, 2021.

Riva Pat, Le Boeuf Patrick, Žumer Maja. 2017. *IFLA Library Reference Model. A Conceptual Model for Bibliographic Information.* https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf. Accessed April 6, 2021.

JLIS.it

SAA (Society of American Archivists) - Council Conference Call. 2018. *Annual Report: Standards Committee and Technical Subcommittees, Appendix D*, 30-44, https://www2.archivists.org/sites/all/files/0118-CC-V-F-Standards.pdf. Accessed April 6, 2021.

Tharani Karim. 2021. "Much more than a mere technology: A systematic review of Wikidata in libraries". *The Journal of Academic Librarianship*, Volume 47, Issue 2, March 2021, 102326, https://doi.org/10.1016/j.acalib.2021.102326. Accessed April 13, 2021.

Vrandecic, Denny. 2013. "The rise of Wikidata". *IEEE Intelligent Systems*, 28(4), 90–95. https://dl.acm.org/doi/abs/10.1109/MIS.2013.119. Accessed April 13, 2021.

W3C. 2013. *PROV-O: The PROV Ontology, Recommendation*, April 30, http://www.w3.org/TR/prov-o/, Accessed April 6, 2021.

JLIS.it

Should catalogues wade in open water?*

Paul Gabriele Weston(a)

a) Università degli studi di Pavia, http://orcid.org/0000-0001-9134-2839

ABSTRACT

In recent years, libraries, either on their own or in consortia, have carried out digitisation projects which resulted in establishing criteria to make digital items accessible through the catalogue. Pushing the boundaries of the latter, cataloguers have considered the possibility of providing access to the digital version of a work whenever available in the public domain. Librarians have now started to question whether the catalogue, moving past the idea of being just a citational tool, should open itself to the web as the place where users, thanks to quality data, can gain easy access to freely available digital bibliographic material. This should include digital publishing, as well as DH projects, all of which are based on editions published in printed format.

This scenario urges to find quick policy answers: a. how should features which could act as search keys or filters be adequately described; b. how should flexibility and changeability of digital objects be dealt with; c. how traditional cataloguing procedures should change as a consequence of the number and the peculiarities of these items; d. which criteria should be adopted in marking the new border lines of the library / catalogue mission.

KEYWORDS

Digital resources description; Metadata management; Digital preservation strategy; Professional education; Catalogue mission; Digital resources retrievability.

* To the everlasting memory of Ottavia Calini, who should have discussed her master thesis on these topics at Ca' Foscari University of Venice.

Setting the scene

The difficult relationship between catalogue and digital production has recently been the subject of reflections initiated within committees and study groups, reflections that were then shared with the library community through conferences and seminars, and professional literature. The issues are far from simple to solve, as they see an overlap of technological factors, cataloguing rules, standards and formats, and procedural choices. A thorough impact assessment of the above-mentioned factors on either the structure or the function of the library catalogue would go far beyond the scope of this paper. Nor is it possible here to ascertain whether and to what extent the changes taking place in the cataloguing rules, data coding structures and information retrieval systems comply with the task of representing the elements of the bibliographic information, which underlies the principles of cataloguing.

A passage from Diego Maltese's introduction to Trombone (2018, 11) can be taken as the starting point of these reflections:

"There's a difference between the library catalogue and a data archive. Equipping the semantic Web with a specific and even sophisticated search engine for resources of all kinds is certainly important, but it is not and should not be, in my opinion, among the tasks of the library.[1]"

Maltese's observation is part of his broader discourse on the concepts of what is inside or outside the boundaries of libraries and catalogues and, consequently, the activities of librarians. Is it the task of libraries to provide data for the semantic web? If so, how important should this activity be considered among those carried out by libraries?

Or wouldn't it be better or wiser to direct intellectual, planning and creative efforts towards improving and refining the search tools of the library tradition and to entrust to the web a more or less wide part of indexing and also the retrieval of descriptions of resources or of the resources themselves, if these are electronic resources? All the more so because, as Sardo (2017, 9) puts it: "new players not previously present on the scene of document management burst forcefully and outclass libraries."[2]

Inconsistencies in digital resources cataloguing

The taboo of describing electronic and digital resources in catalogues has been almost absolute for a long time for a wealth of reasons. In the first place, following a scheme that has occurred whenever a new type of material has shown up, doubts arose on whether the catalogue should include this kind of material. Subsequently, however, the cataloguers experienced some uncertainties as to which criteria to adopt to identify the type of record and the type of material, uncertainties also due to the rapid technological changes and the need to distinguish between the new emerging categories of electronic resources.

---

[1] "C'è differenza tra il catalogo di biblioteca e un archivio di dati. Attrezzare il Web semantico di uno specifico e persino sofisticato motore di ricerca di risorse di ogni genere è certamente importante, ma non è e non deve essere, a mio avviso, competenza della biblioteca."

[2] "Altri attori prima non presenti sulla scena della gestione documentale irrompono prepotentemente e surclassano le biblioteche".

JLIS.it

Today there seem to be two categories of resources that run the risk of being underrepresented or, worse, represented unevenly in the catalogues: these are digital reproductions of printed books and digital editions of textual works available for free on the web.

As far as digitisation is concerned, it would appear to be an optimal solution to indicate its existence by adding a note accompanied by a link in the description of the physical item from which it was taken. For the cataloguer, this process takes a few seconds, since it is sufficient to insert the *uri* of the digital equivalent in a note or a specific field. Once the description of the physical resource has been identified, the user is made aware of the existence of a digital reproduction.[3]

Fig. 1. Link to the digital reproduction of a copy from the record of the paper edition (Source: Opac of the Biblioteca nazionale Braidense, Milano, Italy)

The fact that, as a result of its digital acquisition, the reproduction is formally identical to its original source when displayed on the screen, leads us to think that this type of resource can be considered equivalent to a set of photocopies, a microfilm or a microfiche and thus treated in the same way.

Is that true? Should a digital reproduction be considered the equivalent of the printed item from which it has been scanned?

To answer this question, a number of issues should be addressed:

---

[3] In the December 2020 revision of MARC 21 Bibliographic, the use of field 856 (Electronic location and access) is defined as: "Information needed to locate and access an electronic resource. The field may be used in a bibliographic record for a resource when that resource or a subset of it is available electronically. In addition, it may be used to locate and access an electronic version of a non-electronic resource described in the bibliographic record or a related electronic resource. Field 856 is repeated when the location data elements vary (the URL in subfield $u or subfields $a and $d, when used). It is also repeated when more than one access method is used, different portions of the item are available electronically, mirror sites are recorded, different formats/resolutions with different URLs are indicated, and related items are recorded." (Library of Congress. Network development and MARC standards office 2020)

JLIS.it

From the point of view of the "physical" characteristics of the resource, the answer is negative. The analogue resource, like, for example, the paper book, has its own physical characteristics – the number of pages, the size or the weight – and they are not replicated in the digital object. The dimensions in cm, the rendering of the colours or the weight (these last data not included in the catalogue record) in the digital resource are simulated and suggested, respectively adding a ruler to the video images, a colorchecker, or showing the consistency of the book cut to make the idea of its thickness. When, instead, a viewer allows the reader to directly reach a specific page, the operation is the result of the correspondence created between the specific numbered page and the corresponding digital image. The equivalence between the analogue object and the digital object is then artificially reconstructed for the benefit of those who consult it from the screen.

Even from the point of view of descriptive elements, the scanning of a paper object (but it could be a parchment or a clay tablet) produces a new resource with its own characteristics, starting from the name. Only in early days was it thought that naming the digital object after the name of the analogue resource (for example its title) could be an appropriate solution. For years now file naming has been following criteria unrelated to resource identifiers. For the digital object as a resource in itself, not as a substitute for the analogue resource, in addition to a name that is its own, it might be possible to identify a creator, namely the institution responsible for the digitization project, as well as the entity responsible for its material realisation (for example, the firm that carried out the scanning).

Another crucial element for a complete and correct description is the date of realization of the resource. In the case of digitisation, it is highly unlikely that it coincides with that of the analogue resource, copyright issues being among factors which tend to favour scanning of older resources, not to mention cases where digitisation campaigns are part of special preservation projects of very old originals. The gap between the date of creation of the physical object and that of the digital reproduction is therefore substantial.

Fig. 2. (on the left) De Arte Venandi cum avibus. Ms. Pal. Lat. 1071, Biblioteca Apostolica Vaticana. Graz: Akademische Druck- u. Verlagsanstalt, 1969; (on the right) digital reproduction of folio 49 recto (Biblioteca Apostolica Vaticana, scan date: 23.11.2009).

JLIS.it

What users should be entitled to know

These considerations lead us to believe that it is not appropriate to subordinate the existence and the retrievability of the so-called digital reproductions to their analogue counterparts. But this is what happens regularly. Very few catalogues describe the derived digital objects for what they are, that is, sets of images with specific technical characteristics. Yet, nothing would prevent connecting the analogue object to the derived digital object and providing readers with clear instructions. As it is a right of the latter to be able to identify the existence of one or more digitizations starting from the description of the analogue resource, so it must be equally possible to search and filter the digital resources for the characteristics that are their own, such as the date of creation, an element that could affect in a decisive way the quality of the images and the available exploitation devices, or as the technical characteristics of the images (master and derived) that make up the scanned item. Users may be interested, today and even more in the future, to search for objects created in a given period, as part of a specific project or with specific technical characteristics, not as surrogates, but as objects with meanings other than those of the analogue object. The implications of the application of new IT techniques to the processing of data, both for the purpose of managing digital repositories, and in the process of providing navigation clues to the users, are yet to be fully assessed.

Fig. 3. The Miriam and Ira D. Wallach Division of Art, Prints and Photographs: Art & Architecture Collection, The New York Public Library. "Die Ausstellung von Meisterwerken..." New York Public Library Digital Collections. Accessed April 11, 2021. https://digitalcollections.nypl.org/items/510d47e3-84c9-a3d9-e040-e00a18064a99. The richness of the information and the effectiveness of their graphic layout and, at the bottom of the screen, the time-line that highlights to the reader the significant dates in the creation of the work (author's birth and death, paper edition, digital reproduction, consultation)

JLIS.it

The characteristics, moreover, do not always have connotations of invariability: with the passing of time, whereas it is unlikely that the same institution decides to carry out a new scan of the same item, it is entirely possible that some characteristics of the images, in particular those made available on the web, are modified, such as the format of the file, or that its quality is increased, and therefore the weight, in view of higher performance of computers and connections available to users. There are other elements that have a great relevance in terms of accessibility and that are often linked to the display context: just think of the presence or absence of a menu that provides the document structure, or features such as zoom, OCR, image editing tools (contrast, brightness, etc.), possibility of contextual display of different pages, possibility of downloading high-resolution images and plenty more.

The presence or absence of a navigable summary or, better still, the structure of the document with the indication of pages or illustrations, can make it easier or harder to consult the reproduction, especially in the case of books consisting of hundreds of pages. The same applies to those texts that, having been submitted to the OCR, allow to identify the occurrence of a term or part of it within a volume. This functionality, for example, is not reflected in the paper equivalent and is configured as one of the characteristics of digital objects with the greatest impact on the public. In all cases in which more digital resources are available from scanning the same analogue equivalent, it would therefore be very useful to also provide a description of the services available in the different viewers or on the platforms that host these objects. Considering the question from a diachronic point of view, the description of these services is as crucial as it is subject to obsolescence: interfaces, functionalities and software change in accordance to the available technology and accounting for these developments is definitely complex, especially if the updating work is carried out with conventional procedures.

Fig. 4. Busch, Frank. August Graf von Platen-Thomas Mann: Zeichen u. Gefühle. München: Fink, 1987. The digital reproduction of the volume, carried out within the Digi20 ("Digitalisierung der DFG-Sondersammelgebiete") Project can be accessed at the URL: < https://digi20.digitale-sammlungen.de/de/fs1/object/display/bsb00042052_00001.html?leftTab=PER_ent>. The digital processing has allowed the provision of separate access points for different types of data (names of people, places, references to relevant documents), as well as full-text search.

Going back to the description of the digital resource, it is clear that it should be disclosed to readers not only which exact item was digitised, but also whether an identical digital reproduction is available in multiple versions with different features on different platforms. Describing a digital object through its own characteristics, therefore not as a simple substitute for the analog object, would give the opportunity to generate appropriate filters, but also to create more meaningful links between paper and digital resources. However, these are choices that favour the paper resource and that relegate the digital one to a condition of subordination. To make a comparison, it would be like informing of the existence of an anastatic reprint in a note of the description of the ancient book that it reproduces. Both the anastatic reprint and the digitised reproduction "represent" an existing resource supported by a different medium: coated paper instead of parchment and pixels instead of paper.

JLIS.it

Livello bibliografico	Monografia
Tipo documento	Testo
Autore principale	Accademia della Crusca
Titolo	**Vocabolario degli Accademici della Crusca, con tre indici delle voci, locuzioni, e prouerbi latini, e greci, posti per entro l'opera. Con priuilegio del sommo pontefice, del re cattolico, della serenissima Repubblica di Venezia, e degli altri principi, e potentati d'Italia, e fuor d'Italia, della maestà cesarea, del re cristianissimo, e del sereniss. arciduca Alberto**
Pubblicazione	In Venezia : appresso Giouanni Alberti, 1612 (In Venezia : appresso Giouanni Alberti, 1612)
Descrizione fisica	[28], 960, [104] p. ; 2°
Note generali	· Altro colophon a carta 4L4v: In Venezia : appresso Giouanni Alberti, 1611
	· Segnatura: aⁿ bⁿ A-4Lⁿ a-hⁿ iⁿ; frontespizio con calcografia raffigurante l'impresa dell'Accademia della Crusca; testo disposto in colonne.
Impronta	· a,u- ilne dio- cali (3) 1612 (R)
Nomi	· [Autore] Accademia della Crusca
	· [Editore] Alberti, Giovanni
Luogo normalizzato	IT Venezia
Lingua di pubblicazione	ITALIANO
Paese di pubblicazione	ITALIA
Codice identificativo	IT\ICCU\PUVE\002958

☐ FI0098	CFICF	Biblioteca nazionale centrale - Firenze - FI - [consistenza] 2 esemplari - [tipo di digitalizzazione] parziale - copia digitalizzata
☐ RM0267	BVECR	Biblioteca nazionale centrale - Roma - RM - [consistenza] 1 esemplare - [tipo di digitalizzazione] integrale - copia digitalizzata
RM0521	IEITR	Biblioteca dell'Istituto della enciclopedia italiana Giovanni Treccani - Roma - RM - **Disponibilità temporaneamente limitata; informazioni sul sito della biblioteca** - [consistenza] 1 esemplare - [tipo di digitalizzazione] integrale - copia digitalizzata

Livello bibliografico	Monografia
Tipo documento	Testo
Autore principale	Accademia della Crusca
Titolo	**Vocabolario degli Accademici della Crusca : riproduzione anastatica della prima edizione Venezia 1612 / promossa dall'Accademia della Crusca in collaborazione con Era Edizioni**
Edizione	Rist. anast
Pubblicazione	Firenze : [Accademia della Crusca] ; Varese : Era, 2008
Descrizione fisica	[30], 960, [104] p. ; 35 cm + 1 volume + 1 Cd-Rom
Note generali	· Riprod. facs. dell'ed.: In Venezia : appresso Giouanni Alberti, 1612
	· In custodia
	· Ed. speciale f.c. per Ente Cassa di risparmio di Firenze
	· Edizione numerata.
Comprende	· Una lingua, una civilta, il Vocabolario
Nomi	· Accademia della Crusca
Classificazione Dewey	· 453 (21.) LINGUA ITALIANA. DIZIONARI
Lingua di pubblicazione	ITALIANO
Paese di pubblicazione	ITALIA
Codice identificativo	IT\ICCU\LUA\0531190

Fig. 5. a) Record of the 1612 edition of the "Vocabolario della Crusca", taken from the Opac of SBN. The holding records of three scanned copies are shown below. These reproductions, carried out within distinct scanning campaigns, show different features (the digital reproduction of the National Library in Florence includes only four pages; the copy from the National Library in Rome was entirely scanned within Google Books project and can be downloaded in both PDF and ePUB formats; the digital reproduction of the copy belonging to the Istituto dell'Enciclopedia Treccani is made of just 16 pages taken from various parts of the volume, despite the reproduction is declared "complete"; b) Record of the anastatic reproduction of the same edition, also taken from the Opac of SBN

JLIS.it

If, in the future, the number of digitisations increases and if this is to be adequately and independently represented in the catalogue, the data contained in the hosting digital libraries, whenever available, should be used to create autonomous descriptive data, which then should be properly connected to the original resource.

In fact, in most cases, it can be assumed that the user is not interested in examining the reproduction of a specific copy, but rather the edition, so he or she would be happy to consult the digital equivalent of any specimen. In many other cases, however, his/her interest is directed to the text, the content, regardless of the specific edition. To satisfy this large percentage of research, it would be sufficient to point out, that is, to describe, within the catalogue, the existence of a text or a translation of it in one of the major projects offering works that are now outside the copyright, like Project Gutenberg or LiberLiber. And this brings us to the other category of resources that tends to be underrepresented in catalogues, digital editions.

Fig. 6. Two digital editions of Dante Alighieri's Commedia. The first (top) is taken from Biblioteca Italiana, a project aimed at the publication of texts for study purposes; the second (bottom) is taken from LiberLiber, a project aimed at the creation of a public library, which fact explains the variety of formats

JLIS.it

This certainly meritorious activity is currently not carried out by Italian libraries, with some rare exceptions. At the dawn of the internet, numerous projects concerning the description of web resources (mainly important and authoritative sites) were started, and then abandoned for the poor sustainability, for the difficulty of making the selection and for the instability of the *url*s. The reasons not to indicate the existence on the web of digital texts no longer subject to copyright and freely available on the web may be different. The first is that they are still recoverable through the search engines, motivation certainly correct, but that does not consider how significantly more convenient it would be to be able to find such information in the context that is most appropriate to each individual. It is in the catalogue, in fact, that one can legitimately think of finding books and texts and if access is the immediate one guaranteed by online availability, even better. Failure to report may also be due to the fact that they are not perceived as library resources and that it is therefore not appropriate to devote valuable time to their cataloguing, also in view of the fact that it is impossible to guarantee over time the quality of a web resource, as well as its very existence. In principle, both arguments are correct, even if some digital libraries of texts are now projects of such importance as to guarantee quality and persistence in themselves.

```
▼<TEI.2 TEIform="TEI.2">
  ▼<teiHeader>
    ▼<fileDesc>
      ▼<titleStmt>
          <title>Commedia</title>
          <author>Dante Alighieri</author>
        </titleStmt>
        <extent>711 Kb in UTF-8</extent>
      ▼<publicationStmt>
          <publisher>Biblioteca Italiana</publisher>
          <pubPlace>Roma</pubPlace>
          <date>2003</date>
          <idno>bibit000019</idno>
        ▼<availability>
            <p>Questa risorsa digitale è liberamente accessibile per uso personale o scientifico. Ogni uso commerciale è vietato</p>
          </availability>
        </publicationStmt>
      ▼<seriesStmt>
          <title>Collezione BibIt</title>
        </seriesStmt>
      ▼<sourceDesc>
        ▼<bibl>
            <title>Le opere</title>
            <title type="part">La Commedia secondo l'antica vulgata</title>
            <author>Alighieri, Dante</author>
            <editor id="ed">Societa dantesca italiana</editor>
            <editor id="ed2">Petrocchi, Giorgio</editor>
            <publisher>Mondadori ; [poi] Le Lettere</publisher>
            <pubPlace>Milano ; [poi] Firenze</pubPlace>
            <date>1994</date>
            <note>Edizione nazionale</note>
          </bibl>
        </sourceDesc>
      </fileDesc>
    ▼<encodingDesc>
      ▼<samplingDecl>
          <p>Tutti i materiali paratestuali della fonte cartacea non riconducibili alla responsabilità dell'autore dell'opera sono stati soppressi nella versione digitale</p>
        </samplingDecl>
      ▼<editorialDecl>
        ▼<correction method="silent" status="medium">
            <p>Livello medio: controllo a video con collazione con edizione di riferimento</p>
          </correction>
        ▼<quotation form="data" marks="all">
            <p>I simboli di citazione e di discorso diretto presenti sulla fonte cartacea sono stati rappresentati sulla versione digitale</p>
          </quotation>
        ▼<hyphenation eol="none">
            <p>I trattini di sillabazione a fine riga sono stati soppressi e le parole ricomposte</p>
          </hyphenation>
        </editorialDecl>
```

Fig. 7. The teiHeader of the digital edition of a work contains information that should be made clear in the record of the digital edition itself. In the example, taken from the edition of Dante's Comedy published in Biblioteca Italiana, information is provided on the differences between the paper edition used as the source and the digital edition

As far as e-books and digital publishing are concerned, we should consider the fact that, irrespective of the lack of funds, the sense of national bibliography has disappeared and therefore the preservation for the future consultation of the literary and artistic production of the country is no longer protected.

JLIS.it

Other resources missing

There is a third category of resources which, apart from a few exceptions, are non-existent in catalogues (in particular in large catalogues): these are online resources which libraries acquire not indefinitely, as the single e-book purchased from the publisher's website, but through annual subscriptions. These are the tens of thousands of databases, electronic periodicals and e-books on which libraries now invest the largest part of their budget. At national level, finding out who has access to a database requires knowledge of the Italian library landscape and, in some cases, a good network of acquaintances working in the field.

There is, in fact, no national catalogue of these resources and those who carry out research without obtaining results could reasonably assume the resource in question is not available in any library. There are many reasons for this state of affairs. First, access to these resources is, in almost all cases, limited to users of the purchasing institution through IP recognition or user through ID and password. One might ask, therefore, what is the point of signalling the possession of a resource that is then inaccessible to most.

Again, to make an irreverent comparison, the same could be said for some ancient or rare books, whose consultation is restricted to a very limited number of experts and scholars. Why describing them in a catalogue open to all if only a few have actually access to them?

A second reason is the volatility of the possession of these resources: in many cases the subscriptions are of annual duration and there is always the risk, for budget cuts or in consideration of the scarce use of a resource, that the subscription is not renewed. All the more so for those e-books, we sometimes speak of tens of thousands of titles, which are purchased in packages pre-established by suppliers. The content of these packages changes from year to year, thanks to policies that allow libraries to select the most popular titles to make them part of the library collection. In order to give appropriate cataloguing relief to these titles, it is unthinkable to proceed to the exemplary description for copy. Instead, it is necessary to obtain from the supplier the corresponding descriptive data, and then upload them massively in the cataloguing database. This activity, however, requires a verification of the quality of the data of the authorities present, to ensure that the syndetic structure of the catalogue is preserved, but also a certain timing. The data must be loaded and replaced within a tight time frame compared to the actual availability of the package, otherwise the operation will be useless.

Of course, the accessibility clause is also valid for electronic books for those who have a link with the institution or institutions that own them. Apart from this, electronic periodicals, which are often described at the cataloguing level only in ad hoc portals, such as ACNP, deserve a special mention, while the consistency and availability of an online version are reported only in some cases, in 'traditional' catalogues.

The resources mentioned above can be considered as the digital equivalents of classes of materials that have long been part of the libraries' assets and for which established descriptive standards already exist. There are other types of resources that deserve to be equally taken into consideration on the basis of the importance that their description can play with respect to their visibility, availability and preservation. However, since they are not yet among materials commonly treated by libraries, shared cataloguing criteria are either missing or not yet widely adopted.

The first consists of the products of the so-called digital humanities, a field of studies developed

JLIS.it

in recent years and based on an interdisciplinary approach to research in the humanities and the dissemination of cultural content. Critical editions of texts through computer languages, data visualization, computational linguistics, virtual environments and digital storytelling are just some of the many opportunities of applying computer science to humanities, for example through artificial intelligence techniques such as machine learning to analyze big data and text mining to extract information content from textual content, or semantic web technologies, aimed at improving the understanding of what is asked to the search engine, through associations between information and data.

Fig. 8. The digital edition of a work, carried out as part of a Digital Humanities project, represents, for the purposes of the study, even with all the specific characteristics of the digital application, the equivalent of one or more critical essays and as such should be treated in the context of cataloguing to facilitate its knowledge and access (Source: Francesca Tomasi. *Vespasiano da Bisticci, Lettere. A semantic digital edition*. University of Bologna Centro di Risorse per la Ricerca – Multimedia, 2013. http://vespasianodabisticciletters.unibo.it/#)

JLIS.it

The potential of this area is wide: it allows both to discover new fields of investigation hitherto unexplored, and to expand the public potential of users of humanities through digital technologies, now the main means of production and distribution of knowledge in our society. But for this to happen in a profitable way it is necessary that these achievements, which more and more often constitute the final product of research projects variously financed, are given the same attention as to printed publications. Thus, a number of requirements must be met: the use of open access tools, as well as the adoption of the metadata sets necessary to ensure indexation in cataloguing systems, maintenance and re-use in the later stages of research and storage in long-term repositories

A particular application of digital technology to the publication of reproductions of books, manuscripts and other materials is the International Image Interoperability Framework (IIIF) standard, the purposes of which are described in the following manner:

"The IIIF is driven by a community of research, national and state libraries, museums, companies and image repositories committed to providing access to high quality image resources by defining application programming interfaces that provide a standardised method of describing and delivering images over the web, as well as "presentation based metadata" about structured sequences of images. The standard aims to cultivate shared technologies for both client and server to enable interoperability across repositories, and to foster cooperation among scholars."[4]

For its specific features, the availability of digital reproductions implementing the IIIF should be made known to readers when describing the digital version of a work.
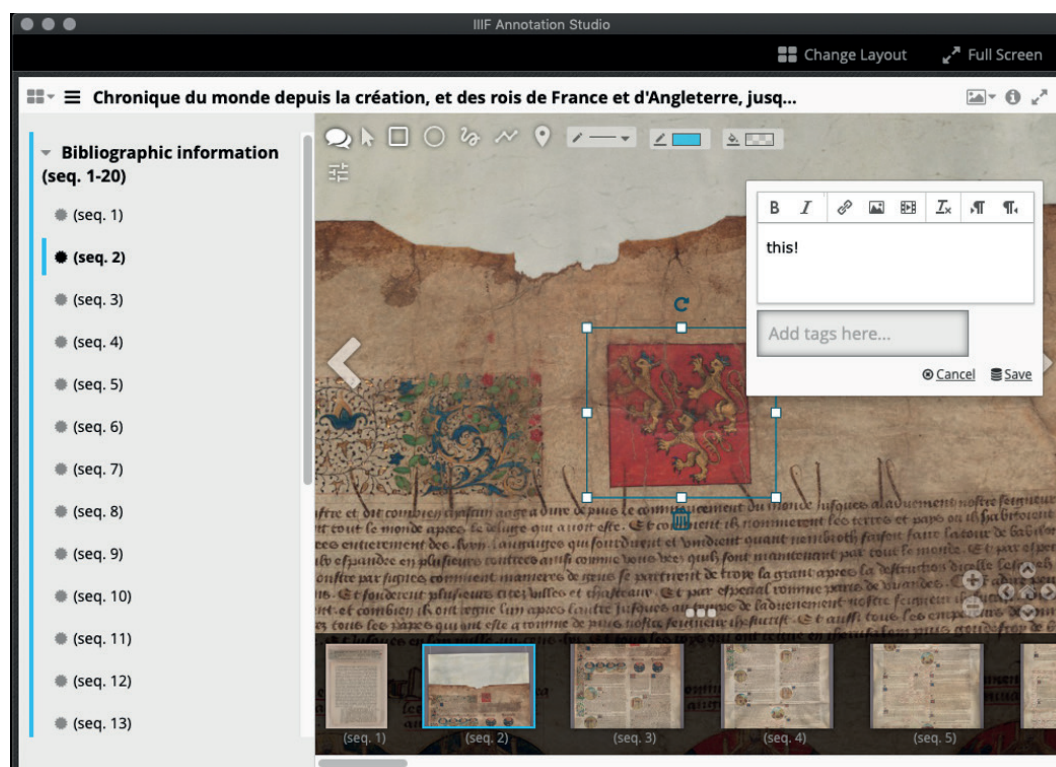
Fig. 9. Users can comment on, transcribe, and draw on image-based resources

---

[4] International Image Interoperability Framework. Enabling Richer Access to the World's Images. https://iiif.io/.

JLIS.it

Other types of resources that should probably be carefully considered are those that have had a considerable boost due to the pandemic. They include, in the first place, the resources created to enable schools and universities to have at their disposal materials useful in supporting teaching and research. Massive open online courses, generally known as MOOCs, online courses aimed at unlimited participation and open access via the Web, were first introduced as early as 2008, and have become over the years a widely researched development in distance education. Aiming at providing open-access features to create virtual environments in which community interactions among students and educators are fostered and supported, MOOCs promoted the reuse and re-mixing of resources such as filmed lectures, readings, data and problems sets. Stemming from this experience, schools and universities have since developed a huge amount of learning objects, "digital self-contained and reusable entities, with a clear educational purpose, with at least three internal and editable components: content, learning activities and elements of context" (Chiappe Laverde, Segovia Cifuentes and Rincón Rodríguez 2007, 8), which require a great deal of investments both in terms of creation and training. To avoid this wealth being dispersed, it is necessary to facilitate their identification, storage and retrieval, through an external information structure consisting of metadata.

Furthermore, universities and professional associations have made extensive use of synchronous and asynchronous streaming sessions, to provide lifelong learning, professional refresher courses and presentations of new products and services. These initiatives too deserve to be preserved, organised, described and made available to the public for future occasions.

The other type of digital resource to consider are podcasts and virtual cultural exhibitions, which in the recent period of the pandemic have enjoyed a large diffusion. Thanks to these initiatives the relationship between people and places of culture and socialization has not weakened, but in some cases has even grown. Libraries, archives, theatres, musical foundations, and opera houses have created a considerable number of products, sometimes revealing a great deal of imagination. Considerable human and professional efforts were required to make all this happen. It would certainly be detrimental not to commit ourselves to preserving and making available this wealth of resources in the future, not just to witness a dramatic event, but as cultural, educational, entertainment, or tourist information materials.

The experience and skills acquired in this last period, together with an interdisciplinary approach commonly referred to as GLAM, and the intersections between the publication formats of the wide variety of classes of digital objects treated, lead to reflect on the way in which digitisation products, the digital libraries, can be reshaped to facilitate their use by users.

It is no coincidence that in the context of the Neustart Kultur Programme, worth almost 1 billion euro, launched by the Federal Government of Germany with the aim of preserving the cultural scene and the cultural infrastructure in the long term, part of the funding has been committed to the programme User-Oriented restructuring of the Deutsche Digitale Bibliothek (DDB). On the assumption that the digitisation projects to which more than five hundred institutions have taken part now give access to 35 million cultural objects in the DDB, 11 million of which are available in digital format, the current programme is aimed at "providing constant free public access to German cultural heritage in digital format still more efficiently. The books, archival materials, photographs, sculptures, paintings, musical works, audio files, films and printed music – in short, the objects – will therefore be linked in such a way that all users of this digital cultural heritage

JLIS.it

will be able to explore it using low-barrier search functions and access it in a user-friendly manner. Cultural education using needs-oriented formats will play a key role in this context. Editorially created content containing participative elements will translate DDB objects and collections into narratives, while collections will be contextualised and presented in formats that can be experienced. The outcome will be a range of services that are easily received and used and that promote interactive participation and orientation amid the diversity of the collections accessible through the DDB."[5]

Changing the paradigm

The presence of digitisation in the catalogues is not a simple cataloguing issue, but it concerns a broader theme, namely the relationship between the world of libraries and the 'outer world', or, to put it another way, the positioning of our activities as librarians. In this perspective, a crucial issue concerns policies regarding the inclusion of resources – other than those owned by the institution –, that are freely available on the web and which ought to be described because of their potential usefulness to library users.

The dilemma is linked to the idea of the Library as an institution, its mission, the role and functions it must perform towards users. And when we talk about functions we do not refer only to clarify, and possibly redefine, the relationship and the services that connect the categories of users that each type of library is called to serve. What needs to be identified and possibly reconfirmed is the role – civil, cultural, recreational, social – which the library carries out in the human context, and which justifies its very existence; a role which should give substance to what David Lankes means by stating that libraries are 'conversations', participatory realities capable of improving our societies.

In this perspective, the aim is to make the library the place to look for works by using its sophisticated search tools, and to save the user the effort to endlessly repeat the search in the chaotic world of the web. Where, then, is the boundary between the library's cataloguing needs, which can be exhausted through now traditional rules and practices, and the possibility of exchanging data with the world of the semantic web at the cost of modifying its structure and also its logic?

The complex story of the development of conceptual functional models is aimed, on the one hand, to make the best use of the architectures of the databases and the way in which software programmes treat and structure the data, and, on the other hand, to allow users to comfortably interact with the catalogue. IFLA LRM, approved at the 2017 IFLA Conference and published shortly after, aims to harmonize, within a new modeling that presents higher abstraction levels, the functional models of the FR family (FRBR, FRAD and FRSAD), to serve as a theoretical reference for metadatation standards, such as, for example, RDA.

For the fact of having been thought of as a versatile tool 'usable' in the semantic web, and consequently based on shared principles and models, independent of the technology used and applicable to any type of medium and resource in any type of cultural institution, RDA raises a number

---

[5] Nutzerorientierte Neustrukturierung der Deutschen Digitalen Bibliothek, https://www.dnb.de/DE/Professionell/ ProjekteKooperationen/Projekte/NeustartKultur/neustartKultur_node.html.

JLIS.it

of issues. In addressing the opportunities offered by RDA, Sardo (2017, 219-225) argues that we are faced with a first step in the direction of a new way of conceiving the activities of cataloguing and of building catalogues that, to deploy its effectiveness, requires the overcoming of a series of significant challenges. First of all, there is the rethinking of the cataloguing data and their organization, which has not yet completely taken place, also because of the huge amount of cataloguing data encoded in ways that are not suitable for the semantic web reality and that cannot always be recoded with automated procedures.

"I wanted librarianship to wake up to the fact that our functional standard was no longer serving us like it should". This was the point that Tennant (2017) intended to make when, in October 2002, he declared in Library Journal that "MARC must die". "I wasn't calling for catalogers to go away. I just wanted something better to work with". That MARC was standing in the libraries' way more than helping them to survive was not that obvious at the time, nor was it an assertion that would pass unnoticed. Fifteen years later, adds Tennant, "no one seems to think it's controversial anymore. The Library of Congress has not only admitted that MARC's days are indeed numbered, they are actively working to develop a linked data replacement. I don't by any means think that we are out of the woods of making this transition yet, and I also believe it will take many years".

The process is, indeed, long and painstaking. The term 'metadata' is now currently used in literature in place of 'cataloguing records', and 'metadata management' has replaced 'cataloguing' in referring to a much wider application context, far beyond the customary library assets. A report produced by Karen Smith-Yoshimura (2020) sheds light on the results of six years of research and discussions within the OCLC Research Library Partners Metadata Managers Focus Group aiming at clarifying changes in metadata due to the awareness that the time of the bibliographic records hosted in silos is rapidly ending, both conceptually and technically. Meanwhile, innovations in librarianship are putting pressure on metadata management practices to move on as the variety of resources for which metadata sets are required is rapidly growing and libraries are even more involved in cross-sectoral projects, both nationally and internationally. The objective to be achieved is plainly summarised in a document produced by the British Library (2019, 2): "Our vision is that by 2023 the Library's collection metadata assets will be unified on a single, sustainable, standard-based infrastructure offering improved options for access, collaboration and open reuse". Expected outcomes of this ambitious process are defined as follows:

- "The complexity of the Library's collection metadata infrastructure will be reduced by convergence on an agreed set of supported standards and systems
- The unified collection metadata infrastructure will offer new access and processing options enabling a greatly improved user experience of Library services
- Efficient, sustainable collection metadata workflows will match the increasing scale and complexity of collection content via implementation of new techniques for record creation and exploitation of external data source". (British Library 2019, 8).

Smith-Yoshimura's report projects these objectives on a much wider scale, to be carried out in countries with very different traditions, organisations, systems of creation and management of data. The question to be addressed as the common starting point of such discussions is: "How do we make the transition to the Next Generation of Metadata happen at the right scale and in a sustainable manner, building an interconnected ecosystem, not a garden of silos?" (Werf 2021).

JLIS.it

If collaboration, agreement upon standard outcomes, reuse of data and ontologies are instrumental in reaching the critical mass necessary to create efficiencies and impact and to generate momentum for the picture to change (Dempsey 2019), at the basis of sustainability is knowledge, and therefore professional education of librarians is crucial. Staff fully aware of the potential of linked data and semantic web technologies, totally confident with the data production process, and reassured that no artificial intelligence, no algorithms are going to undermine human intervention in the production of quality data, are key players in a time of transition. According to literature, substantial investment in the professional training of librarians, as opposed to the simple acquisition of the necessary skills for the execution of mechanical procedures, aimed merely at saving time and reducing costs, looks to be a winning strategy in the long run. Guerrini (2020, 13-14) quite correctly explains the reasons:

> "Cataloging changes perspective and logic by carrying out metadata creation and management, but remains irreplaceable and maintains the distinctive feature of being an activity primarily cultural and, therefore, technical that reflects the ability to analyse and to represent the resources of the bibliographic universe. […] The philosophy of the educational approach to cataloguing cannot be characterised by a dogmatic attitude, but, on the contrary, it requires critical sense and recognition of the editorial and historical complexity of the bibliographic object to be described."[6]

Digital preservation strategies

The question of the relationship between libraries, users and digital resources covers many other aspects that cannot be addressed here, but it certainly cannot be said to be concluded without a reference, however, brief to the issue of preservation, recalling in this regard a thought expressed by Mandillo (2002): "The national collection that is built by law undoubtedly plays a fundamental role in a national policy of freedom of expression and access to information."[7]

The challenge of ensuring that electronic publications are available for future generations is technically complex and resource intensive in terms of both systems and staff. Digital collecting requires new thinking and new processes, in the first place because digital publications, unlike printed material, can be collected once and made available in multiple locations. This gives libraries the opportunity of sharing, together with the collection, the implementation of other management functions, such as description, storage, preservation and delivery. In an ideal situation, the pooling of human, organisational and infrastructural resources should allow to carry out further collection of digital publications, thus increasing the preserved material ratio.

In highly centralised countries, it was the national library that took on responsibility for preserving digital resources deemed to be of interest for cultural purposes. This situation, however, is

---

[6] "La catalogazione cambia prospettiva e logica facendosi metadatazione, ma resta insostituibile e mantiene la caratteristica distintiva di essere un'attività in primis culturale e, quindi, tecnica che rispecchia la capacità di analisi e di rappresentazione delle risorse dell'universo bibliografico. […] La filosofia dell'approccio formativo alla catalogazione non può essere contraddistinta da uno spirito dogmatico, ma, all'opposto, richiede senso critico e riconoscimento della complessità editoriale e storica dell'oggetto bibliografico da descrivere."

[7] "La collezione nazionale che si costruisce per legge gioca indubbiamente un ruolo fondamentale in una politica nazionale di libertà d'espressione e di accesso all'informazione". On the matter of legal deposit in Italy see (Puglisi 2020).

JLIS.it

not very frequent and certainly cannot be the ideal solution in a country such as Italy, where the cultural heritage is dispersed and there are several institutions that have comparable size and history. In addition, almost everywhere there is shortage of staff. Australia[8] and Germany[9] have shown the effectiveness of a strategy based on cooperation between institutions characterized by different nature, size, and field of interest. They have undertaken a long process where nothing is improvised, but is the result of the work of various committees and study groups focused on specific issues.

In this respect, the lesson of Luigi Crocetti is as valuable as it usually is: "Preservation without cooperation is still possible; without cooperation it is not possible to make the library a means of communication and information."[10]

---

[8] See (Lemon, Blinco and Somes 2020).

[9] See (Schrimpf and Tunnat 2019).

[10] "Si può conservare senza cooperare; senza cooperare non si può fare della biblioteca uno strumento di comunicazione e d'informazione".

JLIS.it

References[11]

British Library. 2019. *Foundations for the Future: The British Library's Collection Metadata Strategy 2019-2023*. London: British Library. https://www.bl.uk/bibliographic/pdfs/british-library-collection-metadata-strategy-2019-2023.pdf

Chiappe Laverde, Andrés, Yasbley Segovia Cifuentes, and Helda Yadira Rincón Rodríguez. 2007. "Toward an instructional design model based on learning objects." *Education Technology Research and Development*:671–681. doi:10.1007/s11423-007-9059-0.

Dempsey, Lorcan. 2019. "What Collaboration Means to Me: Library collaboration is hard; effective collaboration is harder." *Collaborative Librarianship* 10 (4, art. 3):227-233. https://digitalcommons.du.edu/collaborativelibrarianship/vol10/iss4/3

Guerrini, Mauro. 2020. *Dalla catalogazione alla metadatazione. Tracce di un percorso.* Prefazione di Barbara B. Tillett. Postfazione di Giovanni Bergamin. Roma: Associazione italiana biblioteche.

Lemon, Barbara, Kerry Blinco, and Brendan Somes. 2020. "Building NED: Open Access to Australia's Digital Documentary Heritage" *Publications* 8 (2):19. doi:10.3390/publications8020019.

Library of Congress. Network development and MARC standards office. 2020. *MARC 21 Format for Bibliographic Data. Update No. 31. 856 – Electronic location and access*. Washington, DC: Library of Congress. < https://www.loc.gov/marc/bibliographic/bd856.html>.

Mandillo, Anna Maria. 2002. "*La nuova legge sul deposito legale: una riforma non solo per le biblioteche*", *AIB notizie* 14 (3):4-7. <https://www.aib.it/aib/editoria/n14/02-03mandillo.htm>.

Puglisi, Paola. 2020. "Deposito legale quattordici anni dopo: come, quando, 'quanto', e perché" *AIB Studi* 60 (3):591-614. doi:10.2426/aibstudi-12477.

Sardo, Lucia. 2017. *La catalogazione: storia, tendenze, problemi aperti.* Milano: Editrice bibliografica.

Schrimpf, Sabine, and Yvonne Tunnat. 2019. "306.2 15 Years of nestor: German Network of Expertise in Digital Preservation (paper Presentation)." OSF. June 20. doi:10.17605/OSF.IO/HA5VN.

Smith-Yoshimura, Karen. 2020. *Transitioning to the Next Generation of Metadata*. Dublin, OH: OCLC Research. doi:10.25333/rqgd-b343.

Tennant, Roy. 2017. "'MARC Must Die' 15 Years On" *Hanging together, the OCLC Research blog*, October 15, 2017. < https://hangingtogether.org/?p=6221>.

Trombone, Antonella. 2018. *Principi di catalogazione e rappresentazione delle entità bibliografiche.* Presentazione di Diego Maltese. Roma: Associazione italiana biblioteche.

Werf, Titia van der. 2021. "Next Generation Metadata… it's getting real!" *Hanging together, the OCLC Research blog*, March 4, 2021. https://hangingtogether.org/?p=8918&utm_campaign=abstracts-6-it&utm_medium=email&utm_source=pardot&utm_content=metadata-opening-plenary-hanging-together-blog-post&utm_term=emea-it-abstracts.

---

[11] Online resources accessed November 11, 2021.

JLIS.it

The National Library of Norway – policies and services

Oddrun Pauline Ohren[a]

a) National Library of Norway

ABSTRACT

The operation of National Library of Norway (NLN) is governed by the Legal Deposit Act of 1989, latest amendment in 2015. By this law, all documents of any type made publicly available in Norway, must be provided to the National Library for registration, preservation and dissemination. According to an added regulation in 2018, NLN may also require the digital version of printed documents, as well as core metadata.

Another important policy document is issued by The Ministry of Culture and The Ministry of Education and Research, outlining a library strategy for the period 2020-2023. While including all types of libraries, the strategy has a strong focus on NLN as a driving force and service provider for the rest of the Norwegian library sector, in mandating NLN to support other libraries in a number of ways, - financially through funding development projects, structurally by way of providing crucial infrastructure and developmentally by conducting our own innovation activities.

The national bibliography forms the backbone for many of the infrastructure services, like the *future Metadata Well*, constituting one single authorized source of metadata for Norwegian libraries, various authority *files*, as well as several *thematic bibliographies*. It also lies at the heart of enabling end users to access the vast collections of digitized material, even much of the IPR-restricted material, obtained through deals with rightsholder associations.

KEYWORDS

National libraries; Governance; Library services; National bibliographies.

JLIS.it

Introduction

A key role of most national libraries is to collect, describe and preserve everything that is published in a particular country. The exact interpretation of "published" and "everything" may vary among countries, – IFLA National Libraries Section expresses its overall goal as "supporting the vital role of national libraries in society as custodians of the worlds' intellectual heritage, providing organisation, preservation of and access to the national imprint in all its forms" (IFLA National Libraries Section 2015). However, during the recent years, openness – both in terms of access to collections and physical space, is seen as an increasingly important value in academic libraries (Anderson et al. 2017, 14, Larsen 2017, 52). This trend is thoroughly embraced by the National Library of Norway (NLN), and is also clearly demanded by the Ministry of Culture. NLN's work on openness first and foremost applies to the collections, and the major premise for that is the mass digitization activities since 2006. Nonetheless, and in spite of a somewhat austere-looking (listed) building, there has also been put strong focus on welcoming students and researchers as well as the general public into the library building, be it for studying, socialising with colleagues and friends or participating in some event.

The following is an account of the NLNs approach to fulfilling its mission.

Governance

The operation of National Library of Norway (NLN) is governed by several policy documents, each with their separate time horizon, from the long term Legal Deposit Act (Norway. Ministry of Culture 2015), via a medium term national Library Strategy (Norway. Ministry of Culture and Norway. Ministry of Education and Research 2019) to the yearly Letter of allocation. The regulatory documents and their influence on NLN's internal strategies and operations are described in more detail below.

The Legal Deposit Act

The Legal Deposit Act of 1989 represents the very "raison dêtre" for NLN, as it defines a stable, very long term basis upon which to construct the national library organisation and its operations. By this law, all publishers, producers or importers of documents made publicly available in Norway, are responsible for providing those documents to the National Library for preservation and dissemination.

The Legal Deposit Act applies to 'any' kind of documents, both physical and digital, on any media – and has done so since 1989. However, through an amendment in 2015 together with an added regulation in 2018 two important changes were stipulated.

Firstly, NLN now may also require the digital files from which the published documents are produced (e.g. pdfs used for printing books), and secondly we may require some core metadata with the deposit. The national library for its part is responsible for creating/enriching the metadata to a level befitting a national bibliography, as well as managing the catalogue and catalogue products.

The Norwegian national bibliography covers several types of materials: Monographs/books, pe-

JLIS.it

riodicals, recorded music published on physical carriers, sheet music, articles in periodicals and resources related to the Sami population in Norway.

These two updates to the law represent great opportunities for streamlining the material flow, and thereby getting the content out to the public faster.

The National library strategy 2020-2023

Another important policy document is *the National strategy for libraries 2020-2023* issued by The Ministry of Culture and The Ministry of Education and Research, outlining a four-year strategy comprising all types of libraries, but with strong focus on NLN as a driving force and service provider for the rest of the library sector. Commonly referred to as *The Library Strategy* for short, and with its emphasis on active dissemination, it mandates NLN to support the other types of libraries in their endeavours, – financially through funding development projects, as well as structurally by way of providing crucial infrastructure to the libraries.

In the words of the strategy document, libraries should develop into "a space for democracy and self-cultivation" (Norway. Ministry of Culture and Norway. Ministry of Education and Research 2019, 3). Moreover, it is points out very strongly that the national library is the government's main instrument and driving force to achieve this, listing a number of duties and tasks for the national library.

The "infrastructure services" to be provided by NLN to other libraries, may be subdivided into 3 groups.

1. Content: The strategy's strong focus on dissemination naturally implies a requirement to provide content, as much as possible digitally, but also physically.
2. Metadata: To administer the content, bibliographical data as well as authority data of good quality are needed.
3. Library tools and guidelines: Part of NLN's duties is also to function as national competence and resource centre for other institutions in the library and cultural sector. This involves keeping up with the development within library science in general and in the bibliographic domain in particular, and provide useful standards, tools and guidance for the whole sector.

Both (digital) content, metadata and authority data should be adapted for machines and humans alike, and the access mechanism must be easy to understand, well documented and readily available for libraries as well as third parties.

Lastly, the guiding principle is that infrastructure services (content, metadata, standards, etc) that NLN provides to the library sector are to be free of charge for end users as well as libraries, – or as cheap as possible.

Summing up, the main message to NLN from the Library strategy is the responsibility to actively disseminate and expand its own collections in various ways on many platforms, to users inside and outside the libraries. Equally important is NLN's obligation to enable other libraries to offer high quality services to their local patrons. To achieve this, NLN shall develop shared, national infrastructure, to make sure that local, often thinly staffed libraries can spend their time and effort to serve their local patrons, not on work that could just as easily be performed centrally at a national level.

JLIS.it

The letter of allocation

This document, issued by the Government after its yearly process of negotiating the national budget, defines NLN's total budget for the year in question, along with any specific areas to focus on, sometimes accompanied by dedicated funding.

The Public Library Act

This law (Norway. Ministry of Culture 2013) stipulates that all municipalities in Norway must offer a public library to its citizens. NLN's role is to enforce the law, in particular the paragraph about competence, instructing each municipality to hire a library manager educated in librarianship.

Content creation and dissemination

The material acquired through legal deposit forms the core of the library's collection, although other material is purchased, in particular documents published abroad which is relevant to Norwegian affairs (Norvegica Extranea). Handling deposited material efficiently is the task which forms the foundation for everything else, and the task that is our sole responsibility. Making the content accessible to users, also involves documenting it in terms of structured metadata, which ultimately forms the national bibliography. Hence, the topic of bibliographical control lies at the heart of the whole process of receiving and processing legal deposit.

Through an extensive digitization project since 2006 – at present about 600 000 books are digitized, practically all the books in our national bibliography. Also, about 60 % of NLN's historical newspaper collection is digitised, and all current newspapers are deposited digitally as well as in paper.

In addition to handling deposited material, NLN also creates content itself, mainly in some way based on the collections. Among these are digital productions like podcasts and streamed events, as well as research-based publications, theme-based bibliographies and re-publications of older literature.

As already mentioned, the national Library strategy focuses very strongly on dissemination: *"… The goal is for libraries to introduce new users to literature and reading, facilitate knowledge dissemination and expand digital collections. The government will implement strategic measures that support libraries and librarians in attracting more users, including those who do not visit libraries."* (Norway. Ministry of Culture and Norway. Ministry of Education and Research 2019, 3)

A large proportion of our physical collection is digitized and therefore – technically – available anywhere through NLN's digital library nb.no[1] and through the main catalogue discovery service Oria. An overall goal is that as much as possible of NLN's content can be accessed throughout Norway. Since much of the material is constrained by copyright, there are legal and financial challenges to be overcome. Consequently, an important part of NLN's dissemi-

---

[1] https://www.nb.no/en/the-national-library-of-norway/

JLIS.it

nation strategy is to negotiate agreements with IPR holders about exposing digitized material still under copyright. At the time of writing, digital books published before 2001 may be read on any device with a Norwegian IP address. The clear message from the Ministry of Culture is that NLN should try to overcome more IPR obstacles, so that books newer that 2001 can be accessed anywhere in the country. Hence new negotiations with the publishers and their organisations will be opened soon.

While great resources are put into maximising the digital content that may be accessed directly by end users, direct availability for end users is not possible for everything. Another agreement with publishers enables patrons of local libraries to access all deposited material from within the walls of their local library, provided it is used for documentation or research. This includes all digitized newspapers and books – also those newer that 2001. Through a special agreement with about 70 running newspapers, any library visitor may read them from 2 weeks after publication onwards.

To provide the same democratic access to physical material for all inhabitants of Norway is no small challenge, Norway's geographical and demographical conditions being as they are. Norway is the 3rd least densely populated country in Europe[2], and the distance from south to north almost spans the whole continental Europe. At the same time, all municipalities are mandated by law [publ act] to offer a public library, however small and however few people live there. Hence, interlibrary lending is by necessity an important element of the library services. To support this, NLN has provided a service called *Biblioteksøk*[3] ('Library search'), a joint discovery&lending service covering the holdings of all Norwegian libraries. Through Biblioteksøk users may reserve books held by any library in the country, and pick it up at their local library.

The NLN Depot library, containing almost a complete set of Norwegian printed monographs, journals, and newspaper microfilms is by far the largest supplier to interlibrary lending. It is built up from legal deposit 'leftovers', transferred material from other libraries and some purchase. The actual lending process is handled by an efficient automatic storage facility. Thus, the Depot library greatly decreases other libraries' burden that interlibrary lending usually represents.

An important part of the Depot Library is the Multilingual Library Collection. Many of the small libraries in Norway, struggle to be able to offer books to their immigrant population. *The multilingual library*[4] is a service for public libraries designed to remedy this to some extent. Its collection comprises literature acquired from a multitude of countries in equally many languages. While being available for all through ordinary interlibrary loan, its main purpose is to support libraries in providing services to their multilingual and multicultural population. Any public or school library may borrow 'mini-collections' or 'book-cases' composed according to their own requirements to be used as their own holdings for a period of up to 6 months.

---

[2] https://www.worldometers.info/population/countries-in-europe-by-population/
[3] https://bibsok.no/
[4] https://dfb.nb.no/multilingual-library

JLIS.it

Research data – The language bank

In order to develop high-quality language technology for Norwegian, big datasets with Norwegian speech and text are needed. Since Norwegian is a very small language in terms of speakers, we cannot rely on others providing such resources. Consequently, NLN Language Bank has as its primary task to provide and organize such data sets. Our resources are aimed at researchers and students, as well as commercial companies developing language technology software. The resource collection comprise among other things, lexical resources like wordnet and dictionaries, corpora of written and spoken language, i.e. large collections of text and speech in machine-readable format. The language bank is NLN's main contribution to the shared European research infrastructure called Clarin[5].

All language resources are available online via the National resource catalogue[6], which also includes resources from other Clarin centres in Norway. The metadata here follows the Clarin-defined framework, in which profiles and subcomponents can be defined, understood and reused across Clarin centres.

Closely associated with the Language bank is Digital Humanities Laboratory, a service supplying scholars, students, and library users with digital tools and methods in their studies, as well as assistance in their use.

The Norwegian national bibliography and other bibliographical services

The NLN is responsible for developing and maintaining an online national bibliography, holding the view that a national bibliography is not merely a tool for information retrieval but is in itself a rich source of insight into a nation's cultural heritage and intellectual production, and as such constitutes valuable research data within many fields of study. While defined by Parent (2008, 10) as "a *current*, *timely*, *comprehensive* and *authoritative* list of all titles published in a country", NLN also includes titles published abroad by Norwegian agents or about Norwegian affairs.

Some parts of the Norwegian national bibliography are still managed in separate legacy databases, but its main portion – along with other, thematic bibliographies – reside as virtual subsets in the main catalogue shared with about 80 other academic libraries.

The whole national bibliography can be accessed through a separate instance of the discovery interface, the discovery service of our main catalogue. The bibliographical data are also freely available via OAI-PMH in MARC 21 format and Dublin Core, to be used for research purposes as well as anything else.

Below is the access page for the Norwegian national bibliography at the search & discovery interface to the catalogue. As shown in the figure, it is subdivided into the subsets of books, serials, musical recordings, sheet music, Sami publications and articles in Norwegian and Nordic publications. The parts that reside in a legacy system are Norwegian registry of serials, Index to articles in Norwegian and Nordic periodicals as well as older parts of the Registry of Norwegian printed sheet music.

---

[5] Common Language Resources and Technology Infrastructure: https://www.clarin.eu/
[6] https://www.nb.no/sprakbanken/en/resource-catalogue/

JLIS.it
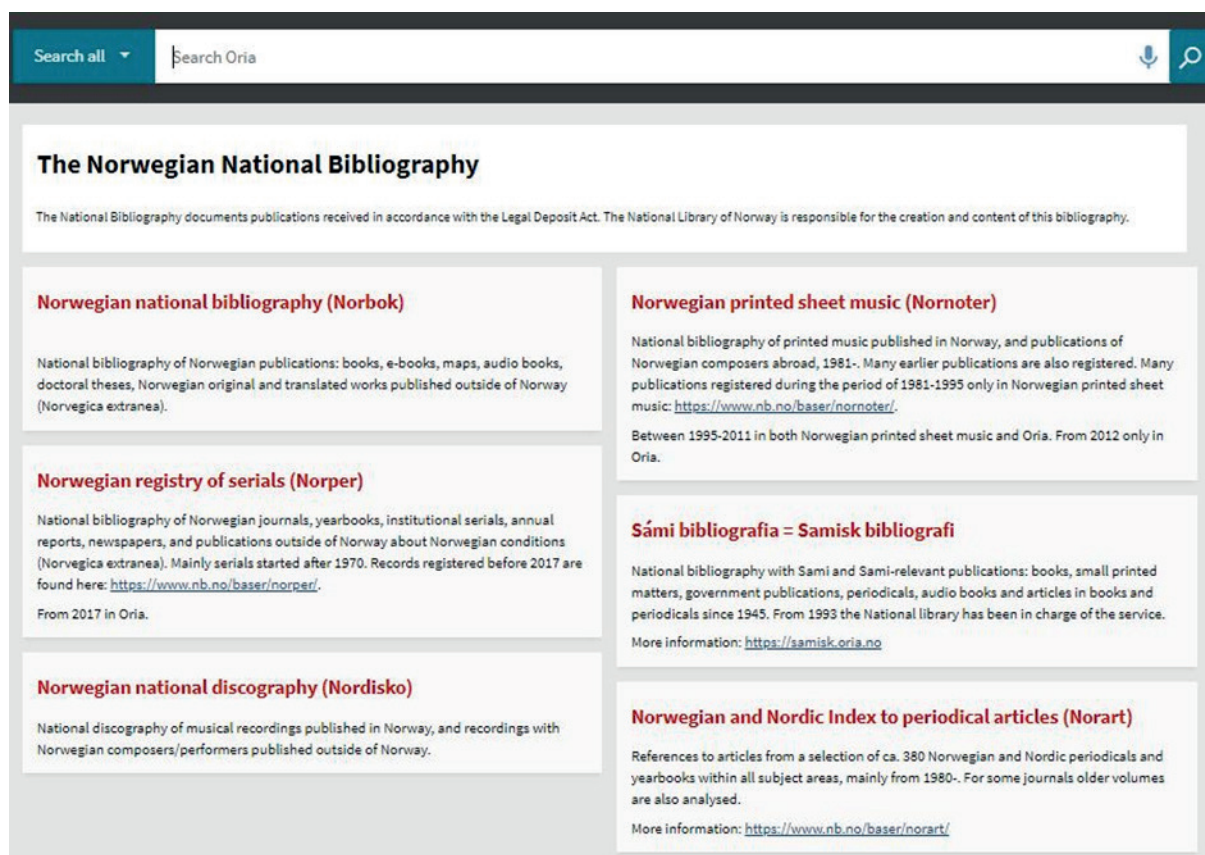
Fig. 1. End user access page for the Norwegian national bibliography

Authorities

Connecting bibliographical data to authorities is an important measure to maintain a certain degree of consistence and reliability in the metadata. So far we have a fairly large authority file for agents (persons and corporations) as well as up-to-date Dewey via WebDewey, a genre form thesaurus published as linked data. A work authority file is in progress, partly through NLN's participation in the library-driven collaboration project Share-VDE[7].

While machine-readable access to agent authorities is provided through a REST API as well as harvesting and download facilities, human users may browse the same 2 million authorities in a dedicated search interface . The set includes all agents referred to by NLN holdings, as well as the holdings of the other academic and special libraries sharing the same catalogue. An increasing number of authorities contain references to other registries, in particular to VIAF[8] and ISNI[9].

---

[7] https://www.share-vde.org/
[8] Virtual International Authority File: http://viaf.org/
[9] ISNI (International Standard Name Identifier): https://isni.org/

JLIS.it

Metadata delivery to the library sector

The current Library strategy mandates NLN to supply free metadata to all libraries. Because prepublication metadata are important sources of acquisition for the libraries, it is important that these metadata be produced as soon as information about a planned publication is available.

NLN's internal processes can at present not guarantee such promptness, hence since 2017, metadata about Norwegian publications (printed book, audiobooks, ebooks and language cources) have been procured from commercial metadata vendors, making sure they follow NLN's cataloguing practice for the national bibliography, and include certain elements particularly wanted by the public libraries, such as subject headings from a certain vocabulary. An important aspect of this is to always authorize agents mentioned in the metadata. Using the API, it is possible to integrate Authority lookup into their own cataloguing system.

This approach will be continued and NLN will assess the feasibility and necessity of expanding to free metadata for other kinds of material such as film and music. Also, libraries have requested access to metadata for foreign material. During the strategy period, the National Library will attempt to find good solutions for including this in its deliveries.

Service overview

Table 1 presents an overview of the services NLN renders as mandated through the Act of legal deposit and the National Library strategy 2020-23, some of which are described in more detail above.

The services are grouped vertically according to type, horizontally according to target group (end user, libraries, third party).

The data services are in principle open for all, there is no difference between any person and a library when it comes to accessing and using our bibliographical data, nor other research data. Content is a different thing, for which copyright is a deciding factor for availability. Negotiating with rightsholders is typically best handled at the national level. So is maintaining cataloguing rules and guidelines, providing core tools like classification system, and providing a website (bibliotekutvikling.no)[10] supporting collaboration and information exchange, as well as serve as a knowledge resource for libraries.

---

[10] https://bibliotekutvikling.no/

JLIS.it

Type of service \ Target group	End user	Libraries	3rd party
Outreach (Competence, governance)	Negotiating agreements with publishers about access	- Metadata standards, guidelines and tools for libraries - Funding development projects - Provide collaboration platform (bibliotekutvikling.no) - Manage Act on public libraries - Negotiating agreements with publishers about access in libraries	- Collaboration forum with system vendors - Interoperability requirements
Content	- The digital library, nb.no - The main catalogue (oria.no) - Events: Digital and on site - Podcasts, social media - Physical material	- Extended access to restricted material for patrons in local libraries - The Multilingual library - The Depot library and Library search	
Data	- Bibliographical data - Authorities: Persons and corporations, Dewey, Genre/form and other vocabularies - Research data: Language resources	- Bibliographical data - Authorities: Persons and corporations, Dewey, Genre/form and other vocabularies - Research data: Language resources	- Bibliographical data - Authorities: Persons and corporations, Dewey, Genre/form and other vocabularies - Research data: Language resources

Table 1. Overview of services from NLN according to type and target groups

As evident from Table 1, producing the national bibliography is only a part of the goals and tasks of the national library, yet bibliographic work forms the basis of most of the other activities and services. For example, many of the events, podcasts and other dissemination activities are directly based on objects in the library's collections. Finding and selecting the right objects in each case requires rich and reliable bibliographic data, describing the objects according to several criteria, like chronology, provenance, topical coverage and physical attributes, among other things.

Challenges and future work

Along with the growing emphasis on dissemination and 'opening up' the library to the general public, comes decreased willingness to spend human resources on cataloguing and related activities, and also stricter demands to justify the usefulness of the particular data elements that are produced. This challenges us to find ways to produce metadata more efficiently with fewer staff, yet maintaining the quality and richness. NLN approaches this from various angles:

Firstly, streamlining the processes handling legal deposit has high priority. The relatively new legislative basis for requiring simultaneous deposit of printing file, printed book and (some) metada-

JLIS.it

ta, offers great opportunities for streamlining the deposit workflow, not least because the need for manual handling of the printed books is greatly reduced. Realising the benefits of this is ongoing work, and is expected to be ready for trial in a few months.

Another potential gain is the possibility for automatic or machine-supported creation of descriptive, based on text analysis of the printing files. Although no concrete action is taken, this will be looked into further down the line.

Machine learning is a type of technology that is starting to gain popularity also in the library universe, especially for subject indexing and named entity recognition, as exemplified in (Suominen 2019). Currently, NLN is in the early stages of experimenting with automatic classification, using the pre-trained BERT model (Horev 2018) for Norwegian.

The Metadata Well vision

One of the most demanding tasks assigned to NLN and its system partner UNIT[11] by the Library strategy is perhaps to establish a so-called 'metadata well': "A further step towards the goal of 'one book, one catalogue entry' will be to create a single authorized source for the metadata – a metadata vault" (Norway. Ministry of Culture and Norway. Ministry of Education and Research 2019, 32). At the time of writing, an RFI document[12] for the Metadata Well is being prepared. Although still very much on the conceptualization and planning stage, the Metadata Well may be thought of as 'an authority file for bibliographical descriptions', in which all metadata produced for the Norwegian National Bibliography are included. So will metadata contributed by other authorized contributors. In its ultimate state of completion, the Metadata Well should contain bibliographic data describing the union of collections in all Norwegian libraries. It is not perceived as a union catalogue, hence no holding data will be included. See also Figure 2 for visualization of the system.
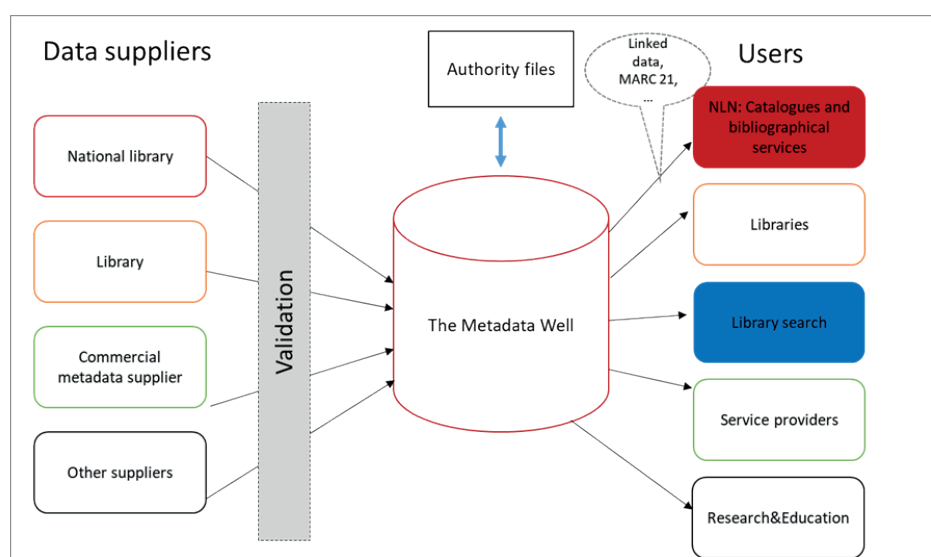
Fig. 2. The Metadata Well in context

---

[11] UNIT Directorate for ICT and joint services in higher education and research (https://www.unit.no/en).

[12] Request for information.

JLIS.it

With the Metadata Well in place, libraries can obtain bibliographical descriptions for their local catalogues, either by referring to it or by copying it to their own catalog, all the while retaining its globally unique identifier in their local data. In their catalogue, they only need to add local information.

Hopefully, this resource will constitute a hub for reuse of metadata between libraries, and as such function as a major source for resource sharing among all types of libraries in Norway, including public, school and academic libraries.

JLIS.it

References

Anderson, Astrid, Cicilie Fagerlid, Håkon Larsen, and Ingerid S. Straume. 2017. "Åpne forskningsbibliotek. Innledende betraktninger." In *Det åpne bibliotek: Forskningsbibliotek i endring*, edited by Astrid Anderson, Cicilie Fagerlid, Håkon Larsen and Ingerid S. Straume. Oslo: Cappelen Damm Akademisk.

Horev, Rani. 2018. "BERT Explained: State of the art language model for NLP." accessed April 15. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.

IFLA National Libraries Section. 2015. Strategic Plan 2015–2017.

Larsen, Håkon. 2017. "Aktivering av nasjonens hukommelse: Nasjonalbiblioteket i offentligheten." In *Det åpne bibliotek: Forskningsbibliotek i endring*, edited by Astrid Anderson, Cicilie Fagerlid, Håkon Larsen and Ingerid S. Straume. Oslo: Cappelen Damm Akademisk.

Norway. Ministry of Culture. 2015. *Lov om avleveringsplikt for allment tilgjengelege dokument (pliktavleveringslova)*. Oslo: National Library of Norway.

Norway. Ministry of Culture. 2013. Lov om folkebibliotek (folkebibliotekloven). Oslo.

Norway. Ministry of Culture, and Norway. Ministry of Education and Research. 2019. A space for democracy and self-cultivation. National strategy for libraries 2020–2023. Oslo.

Parent, Ingrid. 2008. "The Importance of National Bibliographies in the Digital Age." *International cataloguing and bibliographic control : quarterly bulletin of the IFLA UBCIM Programme* 37 (1):9-12.

Suominen, Osma. 2019. "Annif: DIY automated subject indexing using multiple algorithms." *LIBER quarterly* 29 (1):1-25. doi: 10.18352/lq.10285.

JLIS.it

The Italian National Bibliography today

Paolo Wos Bellini[a]

a) Biblioteca nazionale centrale di Firenze, http://orcid.org/0000-0002-5439-1364

ABSTRACT

The statistics on the records produced for the Italian National Bibliography (BNI) in the last decade evidence a stable development with a growing trend. In the face of that, there has been a decrease of human resources never seen before in the history of the National Central Library of Florence (BNCF), that has drastically reduced the editorial staff of BNI to few units. The institutional tasks of BNCF, provided by law, have not changed though. Among these is the archival function for the Italian bibliographic production and its representation through adequate cataloguing and bibliographic instruments.

Therefore, in order to either maintain constant or increase BNI from a quantity point of view, by preserving its quality, some variations of a technical and organizational-managerial nature have been recently implemented, pursuing the following:

1. Rapidity of cataloguing; 2. Full implementation of the recommendations of the Central Institute for the General Catalogue of the Italian Libraries and for the Bibliographic Information (ICCU) as regards, for instance, cataloguing regulations, use of the codes of bibliographic qualification, creation and management of the authority files regarding personal names and uniform title; 3. Constant attention to the role of BNCF in the cooperation within the National Library Service (SBN).

There are plans to intervene in some critical difficulties that still remain. The emergency due to Covid-19 pandemic has imposed a de facto meaningful and unpredictable reorganization of the work management (from the legal deposit to BNI and to the BNCF catalogue) through methods that could certainly be maintained even in future.

KEYWORDS

National bibliographies; Italy; National Central Library of Florence.

JLIS.it

1. Introduction and context

Due to an unprecedented reduction in resources, for years, Italian State libraries have been affected by a deep crisis. This also applies to the National Central Library of Florence in its many branches, including of course the sector where the Italian National Bibliography (BNI) is developed. The group of cataloguers that process BNI is by now reduced to a few staff units. These people, while being indeed few, are very competent and skilled, thanks to a very long professional traineeship carried out in close cooperation with those who made the history of the library science in our Country, such as famous librarians, who, even if retired, often continue to offer, either directly or indirectly, their collaboration. The knowledge of this little group of highly motivated persons draws, day after day, a new vital sap from the job itself, by comparing them in a challenging way with living and always changing reality of the publishing production that these people are called to represent so as to benefit the national and international community of the users.

Indeed, despite the undoubtedly dramatic and even incredible staffing situation the institutional tasks of BNCF, determined by law, have not changed. Such are the tasks that characterize the national libraries and, among them, is thus the establishment and the maintenance of the national bibliographic production's archive. An additional task is also the representation of the national bibliographic production through appropriate cataloguing and bibliographic tools, something which no country wants to renounce or to independently administer through its own national bibliographic agency, that is supported everywhere with adequate human and financial resources.

2. Timeliness

The aspect to which we pay more attention is the timeliness of the cataloguing of the books that the publishers send to the legal deposit office and that are selected, in order to be included into BNI. The promptness of the cataloguing plays a decisive role on the users' fulfilment (whichever group they belong to). This is why it has an absolute priority; yet, among the reasons why it is object of particular interest, is also the need not to create too many bottlenecks and, above all, arrears, within the circulation of the books processed in the library.

Cataloguing at BNCF, and, therefore, also at BNI, is directly implemented through the net of the National Library Service (SBN) by using the software SBNweb (https://opac.sbn.it/opacsbn/opac/iccu/free.jsp). BNCF is one of the 6590 libraries that participate in the SBN net, and so the records processed for the catalogue of both BNCF and BNI also feed the catalogue of SBN and are immediately visible to other libraries of the net as well as, after a short time, in the OPAC of SBN. The records processed for BNI are visible within 24 hours even on OPAC of BNCF, which, in its new version, has been issued only a few months ago (http://opac.bncf.firenze.sbn.it/bncf-prod/).

The selection of the books for BNI in the premises of the legal deposit is carried out once or twice a week. After the selection, the volumes are moved from the legal deposit to the offices of BNI in order for them to be catalogued; normally this happens in a very short time. The availability on the online catalogue of the selected items for BNI is therefore almost immediate. It's worth noting that the bibliographic items created *ex-novo* by BNCF within the net SBN are numerous, which witnesses how quick the book cataloguing on behalf of BNCF is. Just to give an idea, consider that the records of new creation on behalf of BNCF are currently about 784,000 out of 18,320,000

JLIS.it

present in SBN. These items are not represented by the bibliographic records of the documents stored in both SBN and BNCF's libraries, which are obviously many more (since BNCF receives everything that is published), but only by the items that BNCF has created first, compared to other libraries of the net.

The important thing is that within about ten days the books selected for BNI are catalogued and available on OPAC BNCF and on OPAC SBN. There are no arrears.

Yet, in order that the books described by BNI, already visible on OPAC, may show even the semantic contents they were assigned by using the Decimal Dewey Classification and the *Nuovo soggettario*, more time is needed. Indeed, as is well known, the subject indexing is a particularly onerous activity, even if no international bibliographic agency renounces it.

The delay, with which the information of semantic genre of the books catalogued in BNI is shown, is currently equal to a few months, but it is planned to eliminate it shortly with a specific project.

3. Coverage

The other aspect that has a huge importance for every national bibliography and, of course, for the users is thoroughness, that is to say, the quantity of books included every year in the bibliography in relation to the publishing production of the Country.

Books published in the current year and those published up to two years earlier are catalogued within BNI, so in the BNI year 2020 books published in 2020, 2019 and 2018 are included.

Books, which, due to their characteristics, should be indicated in BNI but are not sent by publishers to the legal deposit within said terms, are catalogued by the cataloguing office of BNCF. That means that a special series of BNI files for the related publications does not exist.

In the last five years the number of bibliographic items issued for BNI, referring only to monographs, has been in average about 12.000 per year. This do not include dissertations, periodicals, printed music and other material.

The basic question is how to assess the production of BNI in terms of completeness with respect to the Italian book production. In other words, how representative BNI is as for that. Yet, the concept of bibliographic coverage is therefore not so strict and has to be carefully assessed, in relation to the different typologies of publishing production.

A way to assess these data is to compare them with the statistical data on the Italian publishing production. According to the ISTAT data on the book works published in Italy, in 2018 the Italian publishing houses have published 75,758 titles (https://www.istat.it/it/dati-analisi-e-prodotti/banche-dati/statbase).

This is the number of books to refer to, but, actually, the publications excluded from BNI have always been numerous. In fact, BNI is a selective bibliography. There are several kinds of publications that normally are not reported in BNI. Such publications are of course catalogued by BNCF but only in the catalogue of BNCF (and they are therefore present even on SBN) or they are handled by sets.

The complete list of such publications, not included in BNI, is the following:

1. Official publications of public administrations and international organizations, unless they have an autonomous monographic character.

JLIS.it

2. Non-official publications of laws, decrees, regulations, work contracts, etc., without any comment or addressed to particular categories of readers. The sole exceptions are certain collections of specialized publishers.

3. Publications of parties, unions, chambers of commerce, cultural and religious associations, etc. not of general interest;

4. Pastoral letters and other official documents of religious authorities;

5. Minor religious publications;

6. Consumer literature and reissues of romance novels;

7. Publications not addressed to commerce, but disseminated outside the normal channels of sale, in the form of subscriptions, enrolments, etc.

8. Reissues (unless the publication of reference has never been described), pre-prints, *specimina* and the likes.

9. Complimentary books or gifts, if reissues of previous publications, even if presented in different packages;

10. Manuals and texts for nursery schools, primary schools of first and second level;

11. Biographical scripts for limited use, of either occasional or godly character;

12. Almanacs and the likes of limited interest;

13. Patents;

14. Excerptions, even if presented in only one series;

15. Catalogues of trade fairs and shows prevailingly of commercial interest, catalogues of private galleries, house programs, tourist material;

16. Editorial catalogues and antique catalogues for non-historical or scientific purposes;

17. Publications of promotional and commercial nature, unless they represent the sole or main source of information in special fields, such as complete catalogues of stamps, coins, art objects;

18. Printed music, described in the two half-yearly dedicated files;

19. Texts of lessons and similar materials, in case it clearly appears that they are not addressed to the external dissemination;

20. Speeches and interventions connected to particular events, separately published;

21. Cartographical material.

The material not included in BNI due to choices of publishing policies is thus a lot.

Still in relation to 2018, it is possible to select some typologies of publications, and, by referring to the ISTAT data, it is also possible to verify the relevant quantity, just to give a concrete idea of the numbers we are talking about, without going into the details, which, in this case would be too many.

Schoolbooks, for instance, are 9,786 and children's books are 6,440. Thus, the remaining 59,332 publications, which, aside from reissues and reprints (that are however almost always included in the BNI), decrease to 46,718 publications.

In addition, there are multiple publications, like for instance, text books for primary schools (243), cookery books and recipes (396), books classified as 'entertainment', games and sport (1,004), tourist guides (422), adventure books and detective stories (4,328), comics (713) and others of non-specified genre (1,672) for a total of 8,778 publications. We fall to 37,940, a number still much higher than what remains once selected the afore-said listed items.

JLIS.it

Such list is useful only to give an idea of what numbers we are talking about, but many books belonging to such ISTAT categories are actually included in the BNI (for instance, but not *in toto*, cookery books, tourist guides, detective stories, adventure books and many others).

All this to say that, once eliminated all these typologies of materials, both, in part, reissues and reprints, the remaining books still to be catalogued for BNI are actually those that are catalogued, and that the coverage of BNI is good and representative. Of course, according to the historically adopted criteria, which, although selective, are also fully similar to those adopted by the operators of the same sector, working for both national and foreign libraries. Said operators, while making selections of qualitative type on the material to be reported, do not catalogue at the same level than BNI and are not part of the SBN net. They often resort, in a large percentage of cases, to editorial announcements, which is fine, though for different purposes. In any case, it must be declared and quantified otherwise we shall not supply the users with a quality service.

To end up with this important matter, it must be noticed that many other typologies of books, like all those excluded according to the above-mentioned criteria, not recorded within BNI, are catalogued in BNCF. Just to provide an example, in 2019 the office for the cataloguing at BNCF has catalogued about 24,500 monographs that, summed up to almost 13,000 of BNI, bring the total number of catalogued monographs in 2019 to 37,500, which is a number of all respect if compared to the data provided by ISTAT on the annual publishing production in Italy.

4. The human resources

In order to face the more and more serious and chronic lack of staff, some measures have been taken in these last years:

a) Procurement contracts for the cataloguing to external bodies.
The procurement contracts for the cataloguing to external bodies is a largely known and practical solution. This has both positive and negative aspects. On the one side, such solution guarantees both flexibility and the possibility to 'close up the leak' immediately. On the other side, there is the demand of the quality control that involves BNI at a higher level, even if it obviously involves all; it especially involves BNI because of its role as an Italian bibliographic agency and for the fact that all the books indicated in BNI are catalogued in the SBN net at a "super" level (i.e. level 95), including items marked by the higher authority code. Such records cannot be modified but by the central Institute for the General Catalogue (ICCU) of the Ministry of the cultural Heritage. Whereas mistakes are always possible, said records should not contain any: in other words, they have to meet the requirements of greatest authoritativeness (and here we come to the third key term that must characterize a national bibliography in addition to timeliness and thoroughness/coverage). Both enterprises and cooperatives operating in this field ensure a good level of cataloguing, yet, not always the very high level of specialization required for cataloguing in BNI and that implies heavy investment in staff training.
This is why we must pay a special attention to the control of the correctness of the cata-

JLIS.it

loguing in BNI and, above all, to the alignment of the whole staff. While neglecting additional details, I would like instead to underline that it is not sufficient that the institution may control the plan and the total structure of the service if assigned to external collaborators. On such subject I would like to underline how, due to the reduced permanent staff at BNCF, the quality control subtracts resources in an unsustainable measure and poses a first insuperable limit to the quantity of work that is possible to outsource.

The other limit arises from the maximum fee to be provided for, as established by the Procurement Code.

All this basically means that each tender cannot last more than one year and that, after this term, a new procurement procedure must be performed; in addition, it means that the external staff, in part or completely, changes and that the training work for the external collaborators is each time fully frustrated.

The Covid-19 pandemic has added a further difficulty, by decreasing the number of persons who can work in co-presence in the same premises and thus obliging to temporarily suspend the external collaborations and to seek difficult management solutions.

In spite of the recent important re-adaptations of the rooms to be addressed to the storages, logistic issues that affect BNCF, as well as the scarcity of space in which catalogued books are to be stored, are a further obstacle to the research of viable organization solutions.

b) Cooperation with specialized libraries

Still to cope with the lack of personnel, another important initiative recently taken in BNI was to seek the collaboration with other libraries, somehow similar to BNCF, for the cataloguing of the books reported in BNI. At the distance of a few years, since the beginning of this collaboration, it has been possible to draw up a first balance. It is about a very positive and rewarding experience, which is fully a part of the cooperation and collaboration spirit that characterizes SBN itself. The contribution of the institutes that participate in this activity is however changeable and, in its whole, in a very low percentage. In addition, even in this case, it deals with a very demanding coordination and verification job.

To end with the topic concerning the human resources, every effort has been made to adopt measures of both procedural and organizational engineering to face the above-mentioned dramatic shortage of personnel.

5. Relationships with SBN and library cooperation

The collaboration with SBN is a fact of huge and positive importance for both BNCF itself and the other libraries of the net, because the institutions take mutually advantage on the common job by allowing a considerable saving of working time.

Since the cataloguing in BNCF occurs solely through 'books in hand' and in BNI at the highest level of authority provided by SBN, it can scarcely be said how BNCF does a considerable maintenance job for the catalogue of the National Library Service through the revision of the records processed by other libraries and through the interventions on all the components of the network that gives rise to the card itself.

JLIS.it

One of the decisions recently made was to progressively eliminate every discrepancy out of the application of the rules and of the cataloguers' usage between BNI and ICCU, not only to minimize the needs of intervention on the captured records but also in consideration of the positive value which the uniformity of the choices has in a shared catalogue.

The interventions that are carried out on the records 'captured' by SBN for BNI are very many and I do not fear to exaggerate by assuming that on about 80% of the capture cards it is necessary to intervene more or less significantly.

Listed below are the most frequent and meaningful interventions that are sufficient for me to mention:

- *Ex-novo* publication or updating of the authority file of the personal names with reference to all the books catalogued for BNI which are thus many more than the produced cards.
- In the authority file the codes of the Countries and languages are always enhanced (the code of languages is unfortunately absent also in great part of the cards at a level 97) and the fields "Dating", "Information note", "Sources" and "Cataloguer note" are always compiled through all the proper researches, according to what provided by ICCU guidelines, as for the drawing up of the authority files;
- All data reported in the authority files are monitored on the bibliographic directories normally in usage, such as – for instance – the national bibliographies of various Countries, the catalogues of great libraries, biographical dictionaries, encyclopedia, both national and international authority files, etc., and the connection is executed in all provided and applicable cases;
- If available, the code ISNI is added. In this regard, I remark that in ISNI numerous duplicates are present. When two ISNI numbers are attributed to the same entity, it is not simple to decide which one has to be included in the registration of the name that is being processed. Therefore, comparisons on VIAF shall be carried out, and whenever more incidences of the same name, complemented with an ISNI number, are found even in VIAF, the incidence to which more libraries are connected, and/or the most appropriate one, will be chosen;
- A very remarkable chapter is that of the link to the uniform title, which also constitutes a recent entry for BNI as well as the BNCF catalogue. Until last year, such link was created only for the translations into Italian of foreign works or for ancient works. At present, in full compliance with the FRBR scheme, we follow the SBN regulations, which include the creation of a uniform title in any case. The decision to connect a uniform title for all catalogued publications (and subsequently trans-codified into BNI), in compliance with what provided by ICCU and following the Italian Regulations on the cataloguing for authors (Reicat) for the choice, has definitely caused a remarkable burden of work and an extension of the cataloguing times. It is furthermore important to notice that the application of the rule in SBN is often carried out in a mechanical way, so that the uniform title adopted for the Italian publications, not in translations, always replicates the title itself, by often reporting even the complement of the title, something apparently not always correct. Recently, it has not been rare that uniform titles of foreign works translated into Italian were the Italian title itself. This also implies the need to intervene often in the corrections, fusions, additions of links to variable uniform titles, and so on. It would definitely be preferable that

JLIS.it

ICCU organized other courses especially addressed to the registrations of the authorities. Indeed, the application of the guidelines is not certainly something to be executed though automatisms. Suffice it to think of the complexity of the qualifications and other elements used to distinguish identical titles.

- As per the authority files, BNI participates in the ICCU working groups for the authority files of the names and on the uniform titles. Such interventions are executed by single users who intervene on the authority entries, not through the contextual updating of the Hub database, but rather directly operating on the collective catalogue by means of centralized working methods, that is to say within the so-called "direct interface";
- Proceeding with the reporting of the significant interventions that need be done more frequently on the records 'captured' by the SBN net, the link to an institute is created, still in case of anonymous books, or, in case the books have no main liability, if they are present in the title page, in the cover page or in another relevant position. Such is one of the most frequent interventions on the card 'captured' by SBN (because the link is almost never present), and, due to the controls they bring, said interventions are also very onerous;
- The link to an institute is created even in the case of a series with a generic title just published by that institute;
- Unfortunately, in SBN the cases of records without any element of controlled access are more and more frequent, also when created by libraries of primary importance, even when it is absolutely clear that there are either major or secondary liabilities;
- as already mentioned before, if necessary, the notes are added and, in case there are any, both their adequacy and homogeneousness are controlled;
- the codes of bibliographical qualification are indicated in all the events provided by the regulations on SBN cataloguing;
- relator codes for all the names linked to a new entry are always developed;
- as per the genre of the resource, whose indication is optional, in BNI the codes J "Biographies", S "Exhibitions" and Z "Conference proceedings" are indicated;
- The application of the code "Typology of literary text" has been included.
- The form of both the content and the kind of mediation is obviously indicated (which is indeed mandatory);
- the codes of designation of the type of support are indicated, by using the list of the codes MARC21, managed by the Library of Congress, also used in RDA;
- The link to names, even for not previously considered secondary liabilities, is carried out: such as preface authors, postface authors, and so on;
- Lastly thanks to an important innovation, the indices of the BNI subjects, starting from 2015, have been activating links to the terms of both Thesaurus and New Subject Indexing.

According to the above list, the number and quality of the interventions carried out in SBN are actually high and bring to a very meaningful increase of the granularity of the catalogue, as well as to an enhancement of research opportunities.

JLIS.it

6. COVID-19 emergency

The Covid-19 emergency has imposed a radical reorganization of the work. At first, without any prior experience in the sectors of both BNI and cataloguing, the method of remote working has become mandatory for everyone. The access to the necessary devices was made possible in record time by the IT staff of BNCF within 24 hours. The main problem has been the transportation of the books from BNCF to the cataloguers' homes and vice-versa. This required the willingness on behalf of both the workers and their families. Additional difficulties have been, and still are due to the inadequacy of the equipment available for the employees, as well as the inadequacy of the net, the uncomfortable spaces within their homes, the lack of ergonomic equipment and furniture at their disposal.

The lack of all the needed controlling devices (not everything is available in the web), that, at times, makes it necessary to complete the cataloguing on-site. Both bureaucracy and the restrictive interpretation of the regulation are ever enemies: perhaps, it might be unavoidable, but the adopted registration of the withdrawn and returned books takes precious time away.

Certainly, the emergency has compelled the administration to implement this new working method, something they had been talking about for years, without succeeding in going beyond the stages of the mere projects, more or less created *ad personam.* Now we are able to assess accurately its impact on the library organization and on the real lives of the employees, which has been a huge step forward in only a few months.

Among the so many considerations that can be made, I would like to point out how the Covid emergency has imposed verification processes of the very stringent workflows that makes it possible to assess usefully the daily-executed work.

7. Future perspectives and conclusions

There are also criticism and, above all, objectives that we should set, if we had the necessary human resources to reach them.

Among these, I underline the need to optimize the fruition of the BNI product, too hidden within the site of BNCF, and, therefore, complicated to consult. Yet, what is worth it for BNI is also worth it for other national libraries, that is to say the tendency to incorporate and merge with the catalogues of the national libraries that produce them. Although, this is really a primary target and something on which other national libraries work really very well.

Again, BNI has been making its own data available in PDF format since 2012 and, in XML and UNIMARC, since 2015. The availability of BNI even in RDF format is so far a goal to be reached. We should work in order to give again more space, to enhance the contents and even to change the separate series, besides those of monographs: the periodicals, the printed music and the doctoral thesis.

As already said, a goal to reach as soon as possible is to increase the timeliness with which even the information of semantic type is made available, together with the descriptive record of the books. In conclusion, I think that, from what above described, it seems clear that through the adoption of the multiple measures of organization and procedural character, a very small and more and more exiguous, yet skilled, expert and motivated group of people was able to contrast, in my opinion amazingly, a completely adverse situation.

JLIS.it

They made it possible to increase the grade of thoroughness, timeliness, authority of the Italian national bibliography, at the level that has always characterized it and according to absolute adequate standards, as provided by the National Library Service.

Let's not ignore that the situation is by now extremely critical and that what has been possible so far to do will however not be so for much longer.

The critical mass needed to transform the ideas and the encouragements into concrete projects is going to fail, also because the burden of the daily work has become too overwhelming and the situation is too difficult for everybody. Not only for us at BNI but also for the other sectors of BNCF, as well as for the other organizations with which we should collaborate, starting from the other libraries of the net and of the central Institute of the Ministry itself.

The risk is that a huge and unique heritage made of knowledge, experience, skills, inheritance of generations of librarians, will be lost, without the possibility that it be conveyed to those who will follow. It would be an unrecoverable damage for the world of libraries, for culture, for our Country.

JLIS.it

Artificial intelligence, machine learning and bibliographic control. DDC Short Numbers – Towards machine-based classifying

Elisabeth Mödden[(a)]

a) Deutsche Nationalbibliothek, http://orcid.org/0000-0001-6809-3926

ABSTRACT

Digital publications now account for the majority of new accessions at the German National Library each year. Due to this growing number, it has become quite challenging to collect and catalogue these items properly. At the same time, these changes allow for new ways, in which it can use the collections. For a number of years, the DNB has been addressing the question of how subject cataloguing processes can be automated so that bibliographic records can be enriched with metadata as comprehensively and uniformly as possible. In the course of introducing automated subject cataloguing procedures, work is also being done on the automated assignment of Dewey Decimal Classification numbers. For this purpose, a set of abridged DDC numbers based on is being developed. The article sheds light on how artificial intelligence is used in this process. Furthermore, the challenges posed by the development of DDC short numbers and machine-based classification for different scientific subjects will be addressed. Also, it discusses how the DNB deals with the issues of data provenance, data delivery and quality management.

KEYWORDS

Dewey Decimal Classification; DDC Short numbers; Artificial intelligence; Machine-based classification.

JLIS.it

Introduction

In the German National Library (Deutsche Nationalbibliothek / DNB), both verbal and classificatory subject cataloguing are used for subject indexing. In the course of introducing automated subject cataloguing procedures, work is also being done on the automated assignment of Dewey Decimal Classification numbers. For this purpose, a set of abridged DDC numbers based on, but not limited to, the DDC Abridged Edition 15 and hereafter referred to as DDC Short Numbers, is being developed.

First experiences in the automatic assignment of abridged numbers were gained in the field of medicine (DDC 610). Since 2005, medical dissertations have been classified using a set of 140 DDC Short Numbers. Since 2015, these Short Numbers have been assigned automatically by utilizing artificial intelligence. Short Number sets for other DDC areas are currently being developed. It is planned to extend the automatic assignment of Short Numbers to all subjects and to constantly review the process and its results.

Initial situation

Digital publications now account for the majority of new accessions at the German National Library each year, and the number is rising (see figure 1). In 2020, the collections grew by approx. 1 million online publications like e-books and electronic journal articles. Due to this growing number, it has become quite challenging to collect and catalogue these items properly. At the same time, these changes allow for new ways, in which we can use our collections; for example, it is possible to search for and retrieve individual articles.
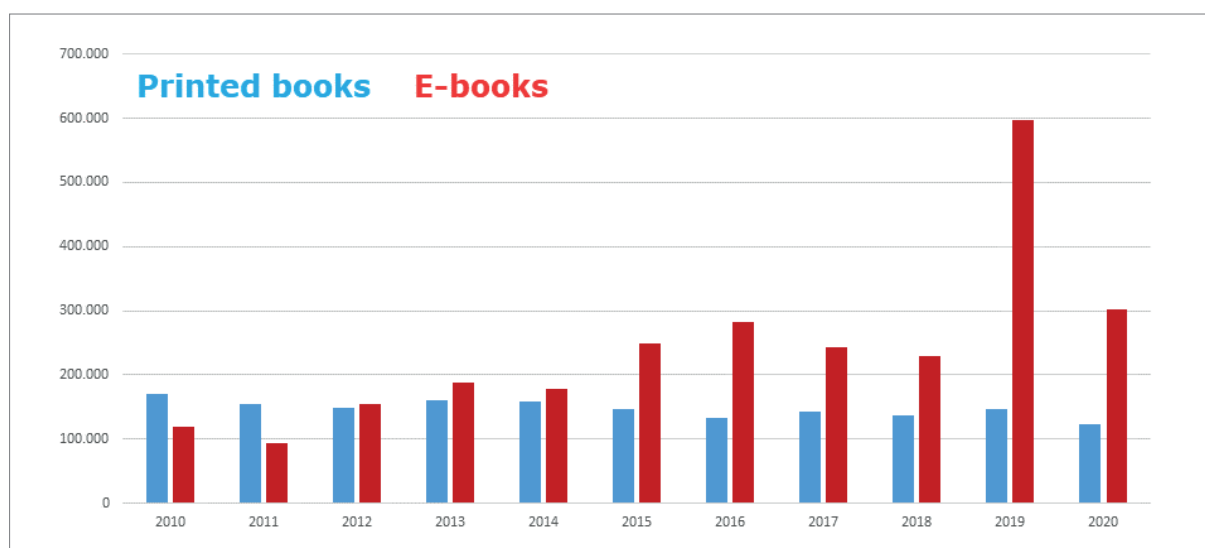
Fig. 1. Increasing amount of publications to be catalogued (e. g. monographs)

Subject cataloguing makes it possible to structure the library's large collections thematically and thereby facilitate the retrieval of publications in these collections. For a number of years, the

JLIS.it

DNB has been addressing the question of how subject cataloguing processes can be automated so that bibliographic records can be enriched with metadata as comprehensively and uniformly as possible – despite new media formats and ever-increasing number of units. Other advantages of automated processes, e.g. the possibility of cataloguing component part of works such as the above-mentioned journal articles both by classification and by assigning subject headings, should be exploited consistently.

Since 2010, the DNB has increasingly been classifying and indexing digital publications using automated procedures rather than intellectual processes (Gömpel, Junger, and Niggemann 2010). In September 2017, the use of machine-based cataloguing procedures was extended to physical publications (Junger and Schwens 2017) ("Cataloguing Media Works" n.d.). In the DNB's Strategic Compass 2025 (Deutsche Nationalbibliothek 2016a) and Strategic Priorities (Deutsche Nationalbibliothek 2016b), the reorganisation of subject cataloguing is addressed as a significant area of activity that will continue to be important during the years to come. This article sheds light on how artificial intelligence is used in this process. Furthermore, the challenges posed by the development of DDC short numbers and machine-based classification for different scientific subjects will be addressed. Also, it discusses how the DNB deals with the issues of data provenance, data delivery and quality management.

Cataloguing methods

Subject cataloguing at the DNB is based on the Series of the Deutsche Nationalbibliografie (German National Bibliography). Every publication catalogued since the bibliographic year 2004 is assigned to one of roughly one hundred subject categories, which are organised in accordance with the Dewey Decimal Classification (DDC) system ("Dewey Decimal Classification (DDC)" n.d.). Beyond that, the publications from the publishers´ book trade provided in Series A are processed intellectually using built numbers from the DDC and subject headings from the Integrated Authority File, the Gemeinsame Normdatei ("Gemeinsame Normdatei (GND)" n.d.).

The development of software applications for subject cataloguing purposes started with the PETRUS project (Schöning-Walter 2010). Machine-based subject category assignment began in 2012, while the automated assignment of subject headings got under way in 2014. Medical publications were first automatically assigned DDC Short Numbers in 2015. At present, work is under way to develop DDC Short Numbers for all subjects.

The DNB employs a support-vector machine for the use in machine-learning processes to facilitate automated classification using DDC Subject Categories and DDC Short Numbers (Mödden and Tomanek 2012). The characteristics of selected text parts and existing metadata are analysed by means of linguistic and statistical methods. During the training phase, the system analyses publications with intellectually assigned Subject Categories and Short Numbers to generate a reference model for all classes of DDC short numbers. When creating this model, it is essential that each class contain sufficient numbers of appropriate learning examples. During the cataloguing process, the system then calculates a statistical measure to determine how closely the content of a new publication matches the patterns learned. As the result of topical classification, the best-matching Subject Categories and Short Numbers are assigned to the publication (see figure 2).

The cataloguing software was created in cooperation with the Freiburg-based company Averbis and is integrated into the DNB's system infrastructure. Machine-based classification has been implemented for texts in German and English.
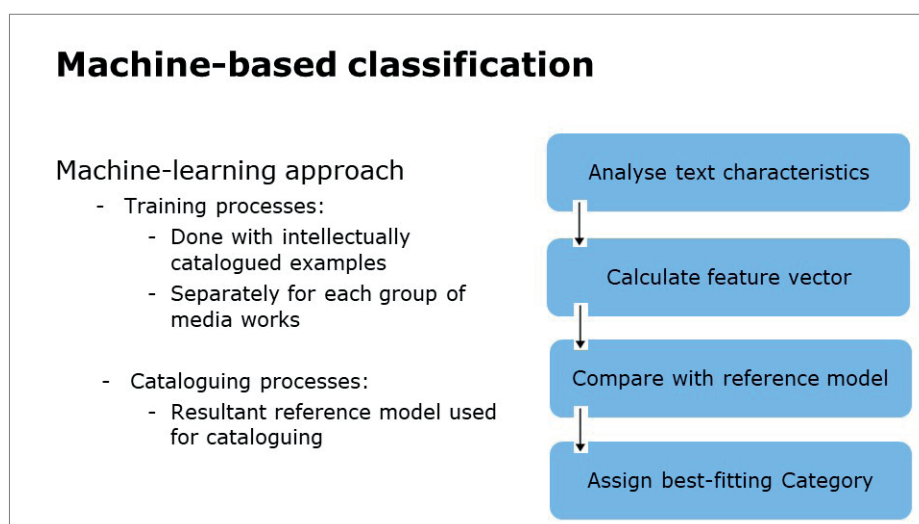
Fig. 2. Processes used in machine-based classification

Workflow

In productive operations, the machine-based cataloguing process (see figure 3) begins automatically at a fixed time every day by sending a list of publications that require first-time processing [1] to a web service. This service retrieves the existing metadata [2] from the cataloguing database (CBS) and the digital full text files or tables of contents [3] from the repository. Before being transmitted to the cataloguing software [4], the storage formats are converted into simple text files and the main language of the publication is determined. Once they have been processed in this way, the results of the analytical process [5] are added to the publication's bibliographic record [6]. Anomalies found during processing are recorded in system files.
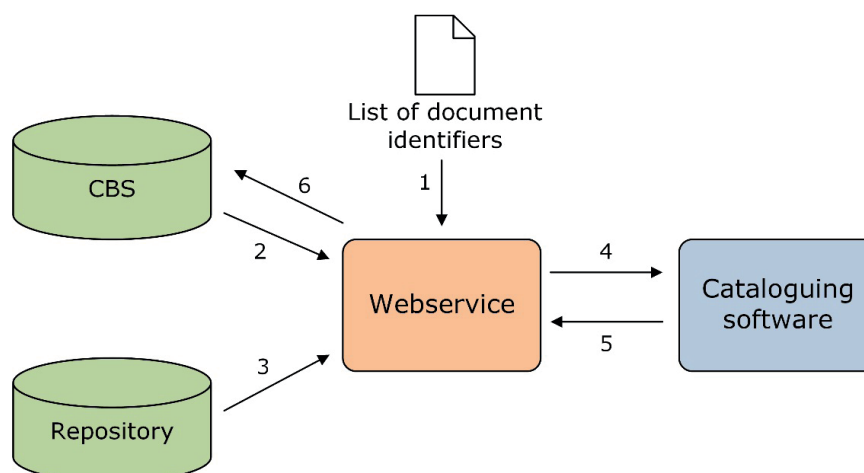
Fig. 3. Technical process used at the DNB for automated cataloguing (productive operation)

JLIS.it

The cataloguing software has various configuration options enabling different types of publications to be processed in different ways. These configurations consist of parameter settings, which have been optimised during test runs. They facilitate the identification of the classification model, for example when assigning Subject Categories. Depending on the features of the publication a certain configuration is set: for instance, digital monographs are processed differently from journal articles, German-language texts are processed differently from English ones, full text files are processed differently from digitised tables of contents.

The software, training corpora and added GND vocabulary undergo regular maintenance and development to ensure that the system as a whole improves constantly. At certain times, digital publications catalogued intellectually are added into the machine-learning processes as new examples. This raises the question whether automated cataloguing processes should be repeated when significant progress is made. In the future, we want to introduce cyclical repetition in order to improve the quality of our machine-generated metadata. We also want to include publication formats that previously were omitted.

Milestones

At the beginning of 2017, the automated cataloguing processes were extended to include journal articles in digital format. The import procedure for e-journals started at the beginning of 2016. Around 675,000 journal articles were integrated into the DNB's collections in 2016 alone. Beginning with journals published by Springer Publishing, the DNB is now enriching individual articles with subject cataloguing metadata. In view of the great number, periodical online publications can only be catalogued economically at this extent by using automated methods.

Another strategic milestone was reached in September 2017 when the automated cataloguing processes were extended to printed monographs in the Deutsche Nationalbibliografie's H Series ("Deutsche Nationalbibliografie" 2019)[1]. Since September 2017 the DNB no longer applies full DDC numbers for this Series. These are to be gradually replaced by DDC Short Numbers. Publications from the publishers' book trade (Series A) will continue to be catalogued intellectually.

In due time, all existing digital resources, for example parallel online editions, tables of contents, abstracts, blurbs and cover texts, will be used for machine-based subject cataloguing of physical media works. At present, publications are catalogued on the basis of digitised tables of contents and the bibliographic metadata that has been supplied. However, since there is less text and a lack of substantial information in some tables of contents the conditions for text analysis are frequently more unfavourable than in the case of online publications. Therefore, the automatically assigned Subject Categories for Series H are all reviewed intellectually.

---

[1] In Series H are university publications: Dissertations and postdoctoral theses from German universities and German-language dissertations and postdoctoral theses from abroad.

JLIS.it

How DDC Short Numbers are selected

Until they are ready for productive use in automatic classificatory indexing, DDC Short Numbers have to pass a multi-stage workflow. Dewey numbers are selected per subject, using the DDC Subject Categories as a guide. This process is accompanied, if necessary, by a comparison with Dewey numbers of DDC's Abridged Edition 15. The next step is to analyse the frequency of occurrence of DDC numbers on the basis of the literature published over the last ten years. Building on this, suitable numbers are selected while numbers with low literature warrant are discarded. This data set is then used for initial technical tests to see how well the selected numbers are working for automatic assignment in the respective subject. Mismatches are analysed and Short Numbers are adjusted in an iterative process. Finally, if the results are convincing, the Short Numbers are put into productive operation and the selection process begins for the next Subject Category. The experts at the Department for Subject Indexing are closely monitoring this iterative process.

Provenance data

The decision to apply automatic processes goes along with the decision to assign the machine-generated metadata to the bibliographic record, to display it in the DNB portal, to use it for retrieval purposes, and to deliver it via the data services. In addition to this, metadata for journal articles is now available in the DNB catalogue and can be obtained through the data services. The DNB's database structure was modified to supply information on the provenance and reliability of the machine-generated metadata. In our database, the machine-generated metadata is recorded together with the date, the configuration name and the confidence value, which is an estimate of the data quality. The machine-generated DDC Short Numbers and subject headings are indicated as such when displayed in the DNB portal (see figure 4).

Link	http://d-nb.info/1211853292
Titel	Keine Auswirkungen des Antibiotikums Norfloxacin auf die Hämodynamik und Rho-Kinase-Expression bei portaler Hypertension im Tiermodell
Person(s)	Bücher-Ollig, Doris Claudia Kristin (Verfasser)
Theses	Dissertation, Rheinische Friedrich-Wilhelms-Universität, 2020
Subject headings	**Norfloxacin*; Tiermodell*; Pfortaderhypertonie*; Leberzirrhose*; Hypertonie*** (*machine generated)
DDC Number	**616.1*** (*machine generated DDC Short Number)
Subject Category	**610 Medizin, Gesundheit*** (*machine generated)

Fig. 4. Title of an automatically catalogued Series O publication displayed in the DNB catalogue with subject headings, DDC Short Number and DDC Subject Category

The data exchange format MARC 21 has also been modified so that standardised information on the provenance of the metadata can be distributed as well.

JLIS.it

Quality and monitoring

Along with daily controls of the process operation, technical checks are carried out by means of sampling. Here, a selection of the publications submitted for automated cataloguing is also classified and assigned subject headings on an intellectual basis. All metadata generated during the cataloguing processes is recorded in the bibliographic database. For display and use in the portal and data services, preference is given to metadata assigned intellectually if available.

For quality management purposes, the quality of the machine-generated classifications is evaluated statistically by comparing automatically and intellectually assigned metadata. Existing metadata for parallel editions is also used for this purpose if applicable. Over the last five years, the DNB has reviewed approximately 18% of the automatically classified online publications in Series O ("Deutsche Nationalbibliografie" 2019). The machine-generated Subject Categories agreed with the intellectually assigned Subject Categories in 76% of cases. This average was actually clearly exceeded in some subject areas, e.g. in law (92% consistency) and medicine (87% consistency). However, machine-based classification does not yet function satisfactorily particularly in the case of subjects on which there is little literary warrant, because there is not enough of the training material required for the learning processes. One such subject for example is the history of South America (DDC Subject Category 980).

There are several issues with machine-based classification of DDC Subject Categories. The main problem is that the machine-assigned DDC Subject Category determines the Short Number. If the Subject Category is wrong, the Short Number will be wrong. Another challenge is posed by the fact that the DDC is continuously updated; even if changes on the broader hierarchy levels do not occur frequently, both changes in the meaning of the class (e.g. change of caption, added or removed major topics) and notational changes such as new or deleted numbers can have an impact on the correct assignment of a Short Number and thus must be taken into account in the process.

Combining machine-based and intellectual cataloguing

Automatic cataloguing procedures are not free from error. Along with imprecise or incorrect assignments, they also generate a bulk of metadata that is not useful for our patrons. The task of quality management is to critically evaluate the error ratio and its effects on the metadata stock in order to adjust the cataloguing processes if necessary. The goal is to achieve a high degree of reliability for the cataloguing data, irrespective of whether it was generated intellectually or automatically. The intellectual and machine-based processes are to be linked more closely in the future. Quality management serves to control and determine which publication forms can be catalogued automatically and which cataloguing services have to be performed intellectually.

Outlook

In 2018, the company Averbis announced a stop to further software developing for machine-based cataloguing. The existing software will be supported only for the next 5 years. Thus, the development of a new machine-based cataloguing system is under way. The target is a new software with a modular structure. This will make it easier to replace individual tools in the future. For this

JLIS.it

purpose, the project "Erschließungsmaschine" – EMa was started. By this, the Averbis software is scheduled to be replaced with a new modular software system by 2022.

Major requirements for the new system are individual modules for text extraction, language recognition, classification, subject indexing, management of text corpora, of terminologies and of notations, etc. After a detailed market study, the Annif toolkit was selected. The National Library of Finland has developed Annif as a tool for machine indexing. The open-source toolkit "uses a combination of existing natural language processing and machine learning tools including Maui, Omikuji, fastText and Gensim. It is multilingual and can support any subject vocabulary (in SKOS or a simple TSV format). It provides a command-line interface, a simple Web UI and a microservice-style REST API." ("Annif – Tool for Automated Subject Indexing" n.d.). For more details, see the very interesting paper by Osma Suominen (Suominen 2019) and the Documentation on GitHub ("GitHub – NatLibFi/Annif:." n.d.). The DNB is very much looking forward to working with Annif, since it is a very promising new tool and is firmly believing that it will pose new opportunities for machine-based classifying and indexing.

In addition, a new, innovative AI (artificial intelligence) project is being launched at DNB. The DNB wants to develop new methods for processing and analysing content and metadata. The new approach should improve the quality of machine-based content indexing in a significant way. Potential AI developments, which are suitable for cataloguing text-based publications, will be investigated, selected, combined and adapted. Research will be conducted to determine which AI methods can be used for machine processing and analysis of natural language texts in order to obtain the most complete and accurate indexing data. The DNB aims for flexibly reusable tools (open-source tools), so that other libraries or institutions with comparable tasks can use these developments as well.

A good database, based on high-quality intellectual indexing by subject experts, is an indispensable prerequisite for the AI project. Therefore, the Department for Subject Indexing will be intensively involved in the development of new procedures. Furthermore, the rules for subject cataloguing should be adapted in such a way as to benefit the combination of both approaches – intellectual and machine-based subject indexing. In the end, the DNB is convinced that high-quality indexing can be achieved by combining intellectual and machine generated classifying and indexing.

JLIS.it

References

"Annif – Tool for Automated Subject Indexing". n.d. Accessed 30 July 2021. http://annif.org/.

"Cataloguing Media Works". n.d. Accessed 29 July 2021. https://www.dnb.de/EN/Professionell/Erschliessen/erschliessen_node.html.

"Deutsche Nationalbibliografie". 2019. https://www.dnb.de/EN/Professionell/Metadatendienste/Metadaten/Nationalbibliografie/nationalbibliografie.html.

Deutsche Nationalbibliothek. 2016a. *2025: Strategic Compass*. Leipzig, Frankfurt, M: Deutsche Nationalbibliothek. https://d-nb.info/1112299556/34.

Deutsche Nationalbibliothek. 2016b. *Strategic Priorities 2017–2020*. Leipzig, Frankfurt, M: Deutsche Nationalbibliothek. https://d-nb.info/1126595101/34.

"Dewey Decimal Classification (DDC)". n.d. December. Accessed 30 July 2021. https://www.dnb.de/EN/Professionell/DDC-Deutsch/ddc-deutsch_node.html.

"DNB_Strategic-Compass-2025_lesesprache_englisch.Pdf". n.d.

"Gemeinsame Normdatei (GND)". n.d. Deutsche Nationalbibliothek. Accessed 30 July 2021. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html.

"GitHub - NatLibFi/Annif: Annif Is a Multi-Algorithm Automated Subject Indexing Tool for Libraries, Archives and Museums. This Repository Is Used for Developing a Production Version of the System, Based on Ideas from the Initial Prototype." n.d. GitHub. Accessed 30 July 2021. https://github.com/NatLibFi/Annif.

Gömpel, Renate, Ulrike Junger, and Elisabeth Niggemann. 2010. "Veränderungen Im Erschließungskonzept Der Deutschen Nationalbibliothek". *Dialog Mit Bibliotheken* 22 (1): 20–22.

Junger, Ulrike, and Ute Schwens. 2017. "Die Inhaltliche Erschließung Des Schriftlichen Kulturellen Erbes Auf Dem Weg in Die Zukunft". *Dialog Mit Bibliotheken* 29 (2): 4–7.

Mödden, Elisabeth, and Katrin Tomanek. 2012. "Maschinelle Sachgruppenvergabe Für Netzpublikationen". *Dialog Mit Bibliotheken* 24 (1): 17–24.

Schöning-Walter, Christa. 2010. "PETRUS – Prozessunterstützende Software Für Die Digitale Deutsche Nationalbibliothek". *Dialog Mit Bibliotheken* 22 (1): 15–19.

Suominen, Osma. 2019. "DIY Automated Subject Indexing Using Multiple Algorithms". *LIBER Quarterly* 29 (1): 1–25. https://doi.org/10.18352/lq.10285.

"The Integrated Authority File (GND)". n.d. December. Accessed 29 July 2021. https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html.

JLIS.it

Annif and Finto AI:
Developing and Implementing
Automated Subject Indexing

Osma Suominen[(a)], Juho Inkinen[(b)], Mona Lehtinen[(c)]

a) National Library of Finland, http://orcid.org/0000-0003-0042-0745
b) National Library of Finland, http://orcid.org/0000-0002-6497-6171
c) National Library of Finland, http://orcid.org/0000-0002-4735-0214

ABSTRACT

Manually indexing documents for subject-based access is a labour-intensive process that can be automated using AI technology. Algorithms for text classification must be trained and tested with examples of indexed documents, which can be obtained from existing bibliographic databases and digital collections.

The National Library of Finland has created Annif, an open source toolkit for automated subject indexing and classification. Annif is multilingual, independent of the indexing vocabulary, and modular. It integrates many text classification algorithms, including Maui, fastText, Omikuji, and a neural network model based on TensorFlow. Best results can often be obtained by combining several algorithms. Many document corpora have been used for training and evaluating Annif. Finding the algorithms and configurations that give the best quality is an ongoing effort.

In May 2020, we launched Finto AI, a service for automated subject indexing based on Annif. It provides a simple Web form for obtaining subject suggestions for text. The functionality is also available as a REST API. Many document repositories and the cataloguing system for electronic publications at the National Library of Finland are using it to integrate semi-automated subject indexing into their metadata workflows. In the future, we are going to extend Annif with more algorithms and new functionality, and to integrate Finto AI with other metadata management workflows.

KEYWORDS

Automated subject indexing; Artificial intelligence; Machine learning; Metadata.

JLIS.it

Introduction

Extensive digitization of paper archives and more active archiving of digital material are creating growing collections of data. Subject indexing, i.e. assigning documents with subjects from a controlled vocabulary, is an important method of organizing collections and improving their discoverability. Traditionally, subject indexing is a manual process performed by human experts, but since manual indexing is a very labour-intensive process, automated and semi-automated methods for subject indexing have been developed since the 1960s (Stevens 1965).

Automating some of the subject indexing processes in Finnish libraries and related institutions has long been a goal of the National Library of Finland for several reasons: to reduce the amount of indexing work, to make the subject indexing more consistent, and to expand subject indexing to collections where traditional manual indexing is not feasible. However, from our perspective, the existing tools and services for automated subject indexing suffer from a number of problems. First, our national languages, Finnish and Swedish, are not well supported by most tools. Second, the tools often rely on their own vocabulary, while we would like to use the General Finnish Ontology YSO[1] (Niininen, Nykyri, and Suominen 2017) as well as other Finnish subject vocabularies. Third, many of the available solutions are commercial services where the customer has little control of the system and is subject to vendor lock-in.

We started the development of Annif[2], our own open source tool for automated subject indexing, in 2017. Three years later, in May 2020, we launched Finto AI – an Annif based automated subject indexing service intended for production use[3]. In this paper, we explain the process of developing Annif, the text classification algorithms it supports, the quality assurance process we use to ensure that the algorithmically produced subject indexing meets expectations, the systems where Annif or Finto AI based automated subject indexing has been deployed, and conclude with some lessons learned.

Development of Annif

The first prototype of Annif was created in 2017, in an experiment to see if it was possible to use freely available metadata from the Finna[4] discovery system to assist in the generation of new metadata (Suominen 2019). After a successful demonstration of the approach, the National Library of Finland decided in 2018 to start the development of a new version of Annif built on a more solid technical foundation and a set of goals and principles:

1. The tool should be multilingual, because in Finnish libraries, there is a need to support at least three languages: the national languages Finnish and Swedish, as well as English.
2. The tool should be independent of the indexing vocabulary; although the General Finnish Ontology is the most commonly used vocabulary in Finnish libraries, other special purpose vocabularies and library classifications such as the Dewey-based Public Library Classification YKL are widely used as well.

---

[1] https://finto.fi/yso/en/
[2] https://annif.org/
[3] https://ai.finto.fi/
[4] https://finna.fi

JLIS.it

3. The tool should support different subject indexing algorithms; a general framework that can accommodate different algorithms was seen as more flexible and adaptable to different situations.

4. The tool should have a command line interface, a web user interface, and a REST API suitable for integration with other systems.

5. The tool should be provided as community oriented open source software; the National Library of Finland advocates for the use of open source software, as part of general openness and transparency goals, and the Skosmos[5] vocabulary publishing software is following a similar open development model.

Based on the above goals, we created a modular architecture for Annif (Figure 1). User interaction is handled either through the command line interface (CLI) or the REST-style API that can be used to integrate Annif with other metadata management systems; the Finto AI web user interface, shown on the left in Figure 1, is an example of such a system. An embedded web user interface that relies on the REST API can also be used for interactive testing.

The *evaluation module* handles the calculation of various evaluation metrics such as precision, recall and F1 score. Annif is configured using a configuration file, handled by the *configuration module*. The *analyzer modules* support tokenization and normalization (stemming or lemmatization) of many languages. Indexing vocabularies, in either SKOS or a simple text format, are handled by the *vocabulary module*. The subject indexing algorithms are implemented as *backends*. The basic unit of configuration is a *project*, which is defined by specifying an indexing vocabulary, language, analyzer, backend, and project- or backend-specific parameters.
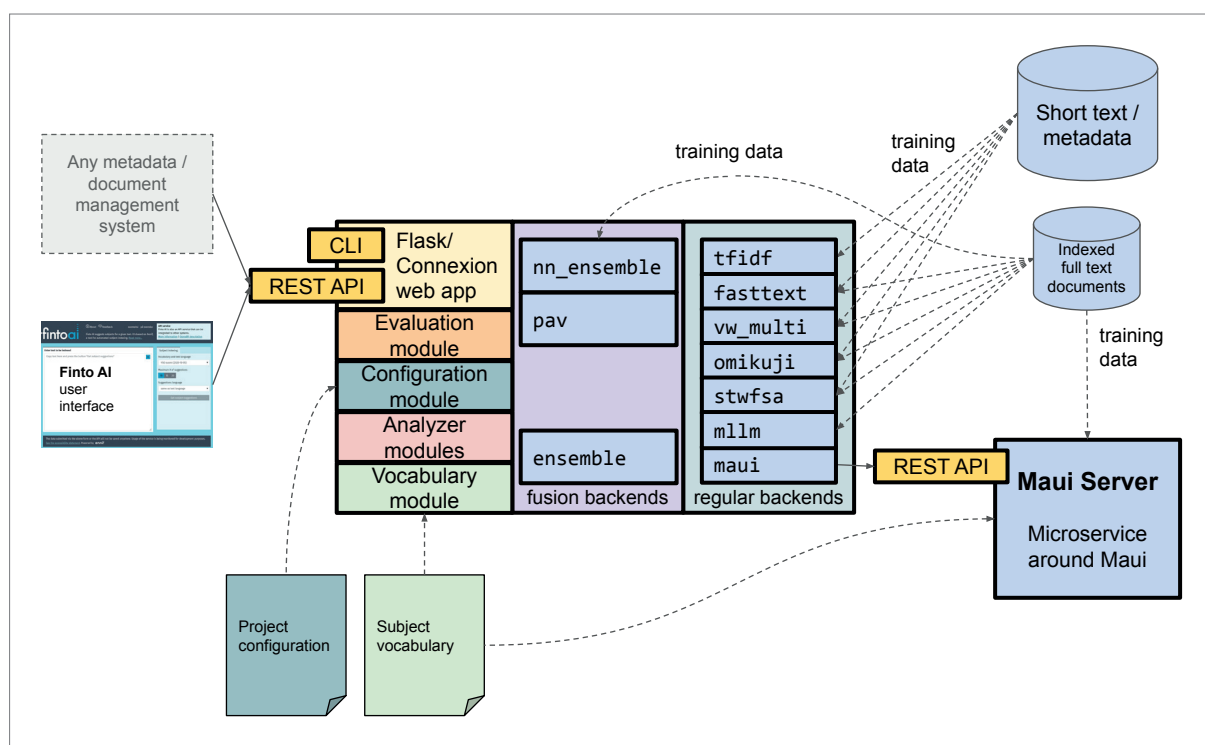
Fig. 1. Modular architecture of Annif

---

5 https://skosmos.org

JLIS.it

Currently all the development of Annif happens on GitHub[6]. Annif is also made available as a Python package[7] and as Docker images[8].

Algorithms in Annif

Annif includes support for several text classification algorithms and thanks to the modular architecture, more can be added over time as backends. Backends can either function as *regular backends* or *fusion backends*. Regular backends work directly on document text and produce a suggestion of possible subjects. The algorithms implemented as regular backends in Annif are based on two main approaches: *lexical approaches* and *associative approaches* (for the distinction, see Toepfer and Seifert 2020). Fusion backends, also called *ensemble* backends, instead use the suggestions from other backends as input and produce a combined suggestion. Backends can thus be stacked and combined in many different ways.

Lexical approaches

In the lexical approach, words within document text are matched with the terms contained in the subject vocabulary. For example, if the vocabulary includes the term *gross national product* and its abbreviation *GNP* (for example as an alternate label for the same concept), then that concept will be suggested as a potential subject for a document containing either the full term or the abbreviation. Since a long document will contain many such matches, lexical algorithms also need to filter and select the most promising candidates; this is typically implemented using heuristics and machine learning.

Maui (Medelyan 2009) is an example of a lexical algorithm, and is supported in Annif by integration with Maui Server[9]. STWFSA (Toepfer and Seifert 2020) is another lexical algorithm supported in Annif by integration with its Python implementation[10]. It has been designed specifically for extracting the maximum information from short text such as metadata records for academic publications. We have created MLLM[11] (Maui-like Lexical Matching), a Python reimplementation of many of the ideas in the Maui algorithm, with some adjustments such as a different string matching method and new heuristics. All the previously mentioned lexical algorithms must be trained with a sample of manually indexed documents.

Associative approaches

In the associative approach, a statistical or machine learning model is trained on a large number (typically hundreds of thousands or more) of manually indexed documents in order to find words

---

JLIS.it

or expressions that correlate with particular subjects. For example, the subject *renewable energy sources* could be correlated with expressions such as "energy", "solar power", "fossil free", "zero carbon", "smart grids" and "battery technology" that appear frequently in documents indexed with that subject, even though not all of them are strictly related to the subject and may not appear at all as terms in the indexing vocabulary. When a well trained associative algorithm is given a new document containing such expressions, it is likely to suggest that it could be about renewable energy sources.

As a baseline method, Annif provides a simple associative backend called TFIDF that calculates a vector representation for each subject based on the words that appear in documents about that subject. When given a new document, the model suggests the most similar subjects for the words in that document, based on vector similarity. fastText (Joulin et al. 2016) is a fast and versatile machine learning algorithm for text classification created at Facebook Research and is supported in Annif by integration through its Python bindings. Vowpal Wabbit (VW) is a general purpose online machine learning framework; Annif supports its algorithms for multi-class and multi-label classification, which are generally best suited for relatively small vocabularies. Finally, Omikuji[12] is a reimplementation of a family of efficient tree-based machine learning algorithms for multi-label classification, including Parabel (Prabhu et al. 2018) and Bonsai (Khandagale, Xiao, and Babbar 2020); it is currently the most versatile and generally best performing associative algorithm in Annif.

Fusion approaches

A fusion approach, i.e. combining different kinds of automated subject indexing algorithms, can be an effective way of improving overall performance (Toepfer and Seifert 2020). Annif provides three fusion backends: a simple ensemble backend, which calculates a weighted average of suggestions from several sources; and two more advanced ensemble backends which require separate training with collections of manually indexed documents. The PAV ensemble (Pool Adjacent Violations) uses *isotonic regression* to estimate probabilities of particular subject suggestions being correct, based on the documents the ensemble has been trained on (see Wilbur and Kim 2014), and combines the estimated probabilities to calculate an overall suggestion. The TensorFlow based neural network ensemble combines the simple averaging method of the simple ensemble with a multi-layer perceptron network that learns how to adjust the combined suggestions so that they best match the manual indexing that the ensemble was trained on.

Quality of automated subject indexing

As we have developed tools and services for automated subject indexing, we have assessed the quality of the automated subject indexing process along the way. According to the framework presented by Golub et al. (2016), the quality of automated subject indexing can be approached from multiple perspectives:

---

[12] https://github.com/tomtung/omikuji

1. Evaluating indexing quality directly through assessment by an evaluator or by comparison with a gold standard.
2. Evaluating indexing quality in the context of an indexing workflow.
3. Evaluating indexing quality indirectly through retrieval performance.

We have so far focused on the first two perspectives, as the retrieval systems affected by the automated subject indexing processes (e.g. Finna) are quite far removed from the subject indexing processes and affected by numerous other factors as well.

API service configurations to evaluate

While we have performed many evaluations of individual algorithms during the development of Annif, the most thorough evaluations have been performed on the combinations of projects, backends, configuration settings, and training data sets that have been provided for public use in the API service for Annif and (since May 2020) Finto AI. The first public API service, after the initial prototype, was published in January 2018, with support for suggesting subjects from the General Finnish Ontology YSO for documents in Finnish, Swedish or English. We set up an ensemble project combining results from three different algorithms for each language. The associative algorithms were trained using metadata extracted from the Finna discovery system, while lexical and ensemble backends were trained on various collections of full text documents. Subsequently we have updated the API service with newer versions of the YSO vocabulary (including YSO Places from January 2020 onwards) and switched the backend algorithms and the ensemble type as new options have been developed. The changes to the API service configurations have been summarized in Table 1.

Date	YSO version	Ensemble type	Backends
2018-01	2017-03 snapshot	Simple ensemble	TFIDF, fastText, Maui
2020-01	2019-03 Cicero	Simple ensemble	Omikuji-Parabel, Omikuji-Bonsai, Maui
2020-03	2020-01 Diotima	Neural network	Omikuji-Parabel, Omikuji-Bonsai, Maui
2020-12	2020-10 snapshot	Neural network	fastText, Omikuji-Bonsai, Maui
2021-04	2021-03 Epikuros	Neural network	fastText, Omikuji-Bonsai, MLLM

Table 1. API service configurations.

Comparison to gold standard

A gold standard is a collection in which each document is assigned a set of subjects that is assumed to be complete and correct (Golub et al. 2016). Once a gold standard has been developed, it is easy to evaluate automated subject indexing methods against it by measuring how well the algorithmic suggestions match the gold standard. However, creating a good quality gold standard takes a significant amount of effort and requires input from many experts. In practice, existing manually indexed documents are often used as a substitute for a properly constructed gold standard, as in

JLIS.it

the evaluation of the Maui algorithm (Medelyan 2009). Such collections are readily available and they enable easy experimentation and comparison of different algorithms, but as the indexing process is susceptible to many kinds of bias, they are best used as ballpark estimates of quality and must be complemented with other types of evaluation.

We have used the following manually indexed corpora for evaluation. The first three include documents in Finnish, Swedish and English, the last two are only in Finnish.

1. JYU theses: Master's and doctoral theses from the University of Jyväskylä (n=7,400) published in the years 2010 to 2017. These are long, in-depth academic documents that cover many disciplines.

2. Electronic deposits: Non-fiction electronic books (n=9832) published between 1998 and 2019 that have been deposited to the National Library of Finland and indexed in the national bibliography Fennica.

3. Book descriptions: Titles and short descriptions of non-fiction books (n=51309) collected from the database of the book distributor Kirjavälitys Oy, covering the time period from approximately 2000 to 2019. The book descriptions were originally created by publishers for marketing purposes. The subject indexing for these works was obtained separately from the national bibliography Fennica.

4. Ask a Librarian: Question and answer pairs from the Ask a Librarian service run by public libraries in Finland. The original database consisted of over 25,000 documents but we extracted the subset with a minimum of 4 subjects per document (n=3,150). These are short, informal questions and answers about many different topics.

5. Satakunnan Kansa: Digital archives of Satakunnan Kansa regional newspaper. The archives consist of over 100,000 unindexed documents. Out of these, a random sample of 50 documents was manually indexed by four librarians working independently.

We split these collections into train, validate and test subsets. Only the test subsets were used as gold standard sets for the evaluation of algorithms. We mainly used the F1@5 metric for the evaluation: that is, the F1 score similarity (harmonic mean of precision and recall) between the manually assigned subjects and the top 5 suggestions of the algorithm. The results are summarized in Figure 2. We can see that the overall F1 scores have generally improved with successive API service configurations. The best F1 scores of around 0.6-0.7 were obtained with Swedish language documents from the JYU theses and electronic deposit collections; however, these measurements are also the least reliable, since due to the small number of Swedish language documents in these collections, we had reused some of the same documents for both training and evaluating the Maui, MLLM and neural network ensemble models. If we exclude the two Swedish language collections with unrealistically good results, we have reached F1 scores ranging between 0.3 and 0.5.

JLIS.it
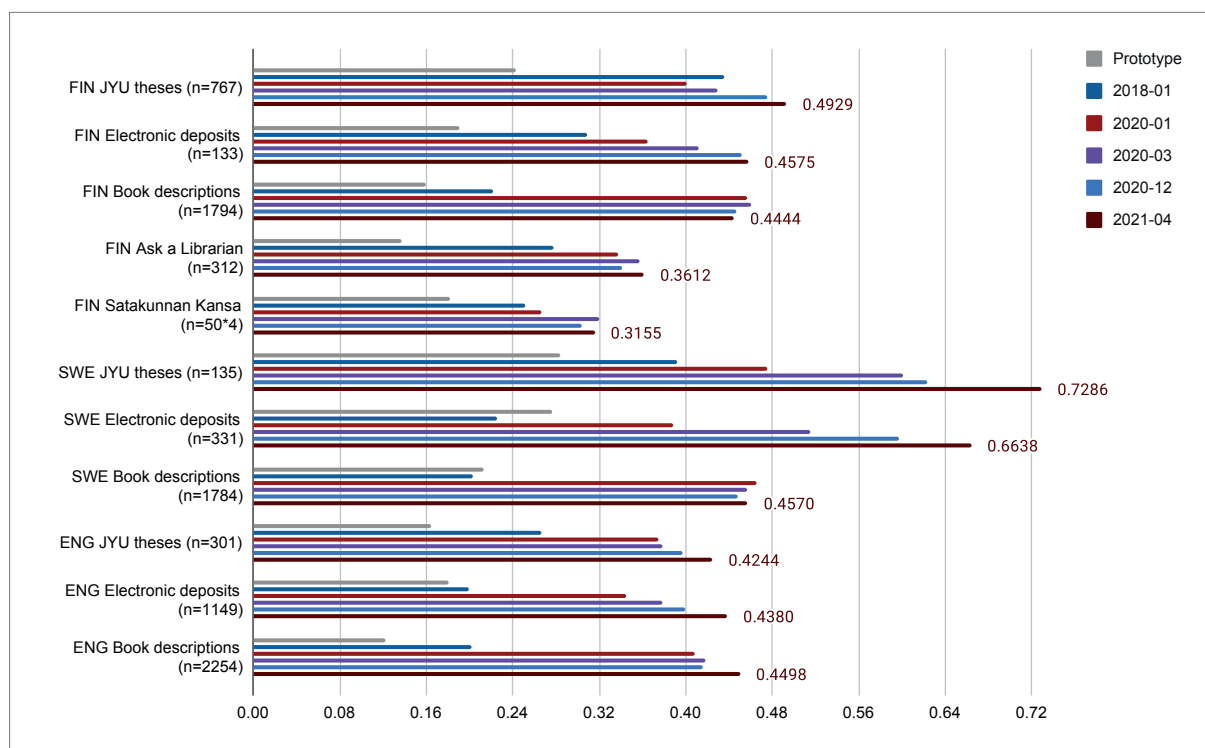
Fig. 2. F1@5 scores for the test collections, by API service configuration. The most recent numeric scores for the 2021-04 API service configuration are also shown

Assessment by evaluators

Having human evaluators assess the suggested subjects is another way to measure the quality of automatic subject indexing. In 2019, we organized a workshop where 48 participants (mainly librarians and informaticians) were given 50 example documents, with on average more than 10 sets of subjects assigned to each document. The indexing had been created either by humans (professional or lay) or by different Annif algorithms, but the participants did not know which was which. The participants used a scale from 1-5 to evaluate the indexing from three viewpoints: overall quality, meaningfulness and coverage. In general the human assigned subjects got higher scores, but the difference wasn't very large. Figure 3 shows the evaluation results. Indexing by the best performing Annif PAV ensemble model usually received a grade of around 3 out of 5, while human indexers scored between 3.5 and 4 out of 5, with professionals performing the best. Annif-assisted semi-automatic indexing landed in between. (Lehtinen, Inkinen, and Suominen 2019)

JLIS.it
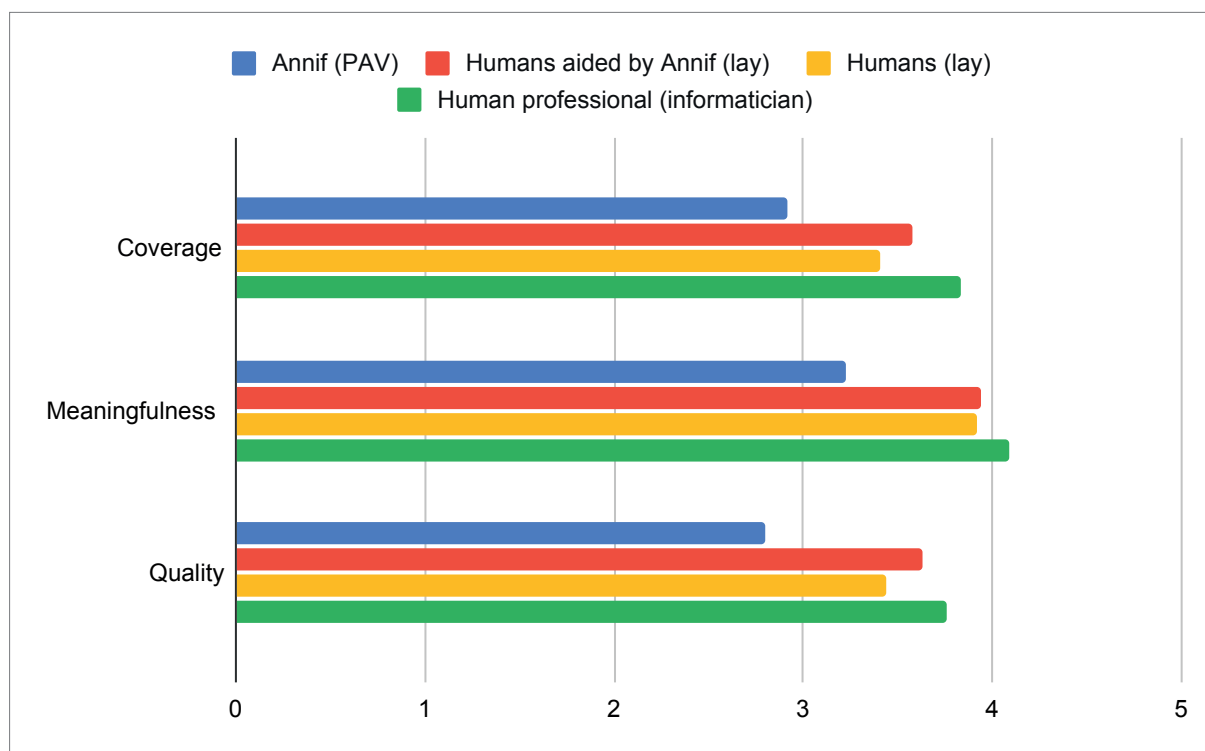
Fig. 3. Quality evaluation of intellectually given and Annif-produced subject indices. Data reproduced from Lehtinen, Inkinen, and Suominen 2019

A similar comparison was performed by the Finnish Public Broadcasting Company Yle. Their tests compared Annif against a commercial document classification service Leiki which they have been using in production for several years. In their results, Annif was rated as slightly better than Leiki for Finnish language documents and as much better for Swedish language documents. The quality of the metadata they used for training Annif might explain the differences between the languages (Suominen and Virtanen 2020; Nikkarinen 2021).

The Research department of the National Library of the Netherlands has also evaluated and used Annif as a part of developing their own larger tool for automated indexing (Haighton and Veldhoen 2020). The German National Library has evaluated Annif as well, comparing it with their current automated indexing system both qualitatively and quantitatively. Seven out of nine Annif's algorithms outperformed the current solution in F1@5 scores. Human evaluators also rated Annif's suggestions as more useful than those of the current system (Uhlmann 2020).

Evaluating in the context of an indexing workflow

We have also evaluated the quality of Annif in the context of the indexing workflow of the JYX[13] institutional repository of the University of Jyväskylä, which was an early adopter of Annif. JYX

---

[13] https://jyx.jyu.fi/

JLIS.it

integrates Annif into its upload form. Students who upload their completed Master's thesis receive suggestions from Annif and can accept or reject the suggestions as well as add their own keywords. Later in the process, informaticians validate the metadata and can make corrections to the subjects. The system saves the original suggestions by Annif as well as the users' choices, so it is possible to keep track of how many of the Annif suggestions are accepted by the student and the final validated subjects. Figure 4 shows the F1 score similarities between the original Annif suggestions and the student-selected and final subjects, over several generations of API service configurations. There was a marked increase in the similarity after the initial prototype; since then, a small increase in similarity to the Annif suggestions can be seen in both the student-selected and final subjects, suggesting that the acceptability of the automated suggestions has increased over time.
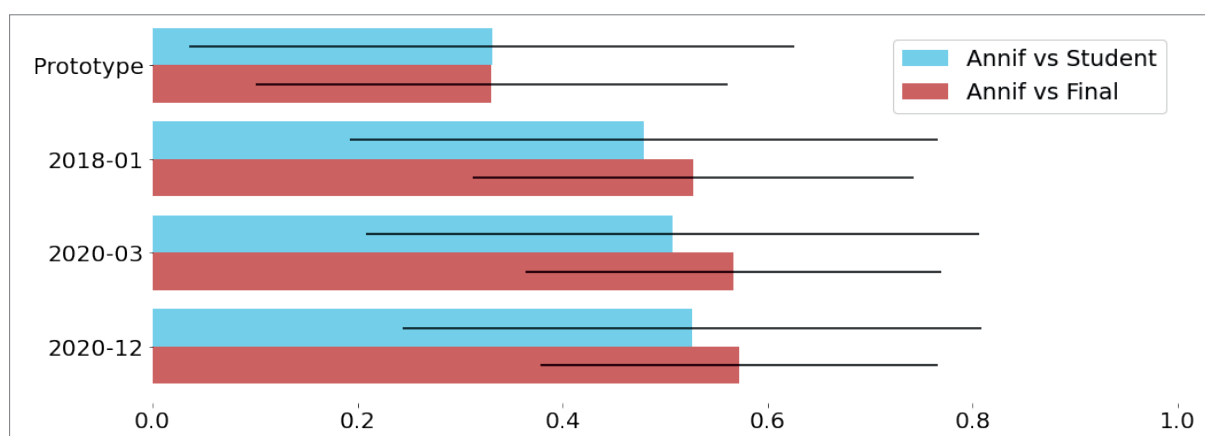
Fig. 4. F1 score similarity between Annif suggestions, student-selected subjects and final subjects in JYX, for the Annif prototype and subsequent API service configurations. Data is missing for the short-lived 2020-01 configuration

Users of the Annif API service and Finto AI

A service for automated subject indexing based on Annif has been existing since 2017 at the annif.org website, but its main purposes have been testing and development. The Finto AI service we launched in May 2020 is intended for production use. The service offers an easy way for introducing automatic subject indexing into information systems, provided that the vocabularies and language support offered by the API service meet local requirements. Some of the systems integrated with Finto AI are shown in Figure 5.

Generally, when the API service is integrated in the indexing workflow of a document repository, the steps in processing a document are:

1. extract the text from the document (typically a PDF file)
2. detect the language of the text (if not already known)
3. send the text to Annif via the *suggest* method of the API; the specific endpoint is chosen based on the text language and the indexing vocabulary
4. display the returned subject suggestions to the user
5. the user selects the subjects to be stored in the document metadata; the user can also add subjects that were not suggested by Annif

JLIS.it
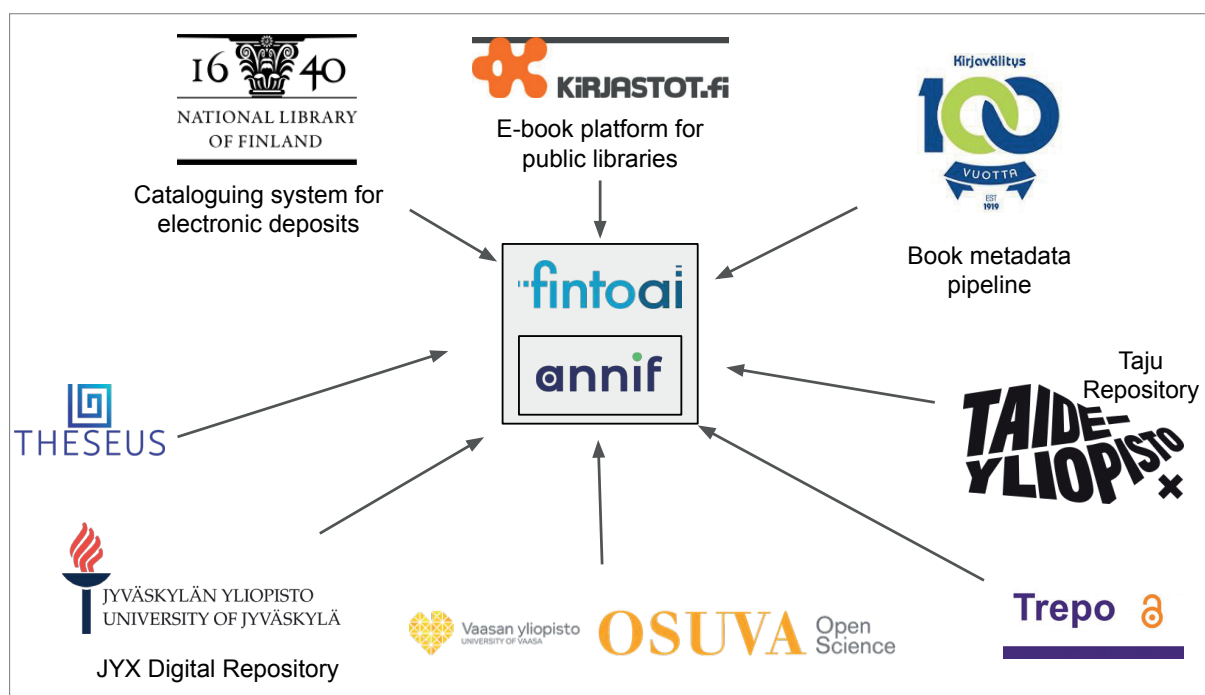
Fig. 5. Institutional users of Finto AI

Institutional repositories

The very first institutional user of semi-automated subject indexing by Annif was the JYX repository of the University of Jyväskylä, which is based on DSpace software. Already in 2017 they integrated the API of the Annif prototype system into their pipeline, which is used by students to upload their Master's or doctoral theses. As explained above, a librarian may correct the student-selected subjects when validating the metadata.

Since 2020, until April 2021 when this article was written, four DSpace based university repositories maintained by the National Library of Finland have started using Finto AI in their uploading pipeline: Osuva[14] (University of Vaasa), Trepo[15] (University of Tampere), Taju[16] (University of Arts) and Theseus[17] (used by many Finnish universities of applied sciences). Their workflow is similar to JYX.

The electronic deposit system at the National Library of Finland

The National Library of Finland maintains an uploading service for individual deposits of electronic publications[18]. The API of Finto AI was integrated in 2020 to the metadata workflow of the

---

[14] https://osuva.uwasa.fi/
[15] https://trepo.tuni.fi/
[16] https://taju.uniarts.fi/
[17] https://www.theseus.fi/
[18] https://luovutuslomake.kansalliskirjasto.fi

JLIS.it

internal deposit repository Varsta. The subject suggestions are not shown to the uploader, but to a library cataloguer who curates the metadata in the Varsta system. The metadata is then stored in the Melinda union catalogue. The publication files are stored in the Varia repository, which can be browsed using the computers within the premises of the National Library of Finland. See Figure 6 for an overview of the pipeline.
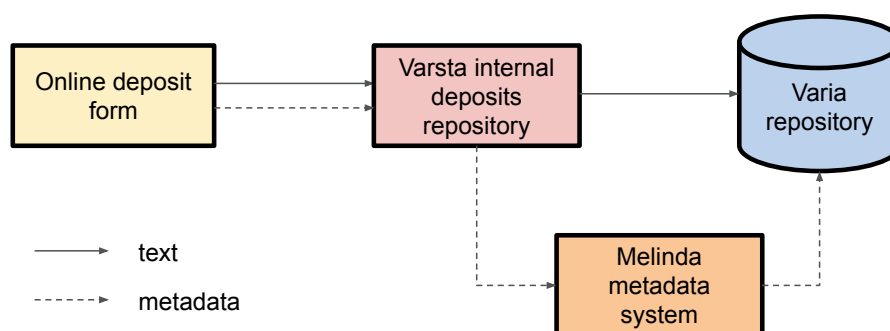
Fig. 6. Data flows for individual electronic publication deposits in the systems of the National Library of Finland

Book distributor Kirjavälitys Oy

Kirjavälitys Oy[19] is a Finnish book distributor that handles book-sale logistics. They receive information about upcoming titles from publishers and produce metadata used by libraries, booksellers and the union catalogue Melinda, which includes the Finnish national bibliography Fennica (see Figure 7). Kirjavälitys has integrated the API of Finto AI in their system to aid in subject indexing of non-fiction books. They use the back-cover description text of books as the input to Finto AI.
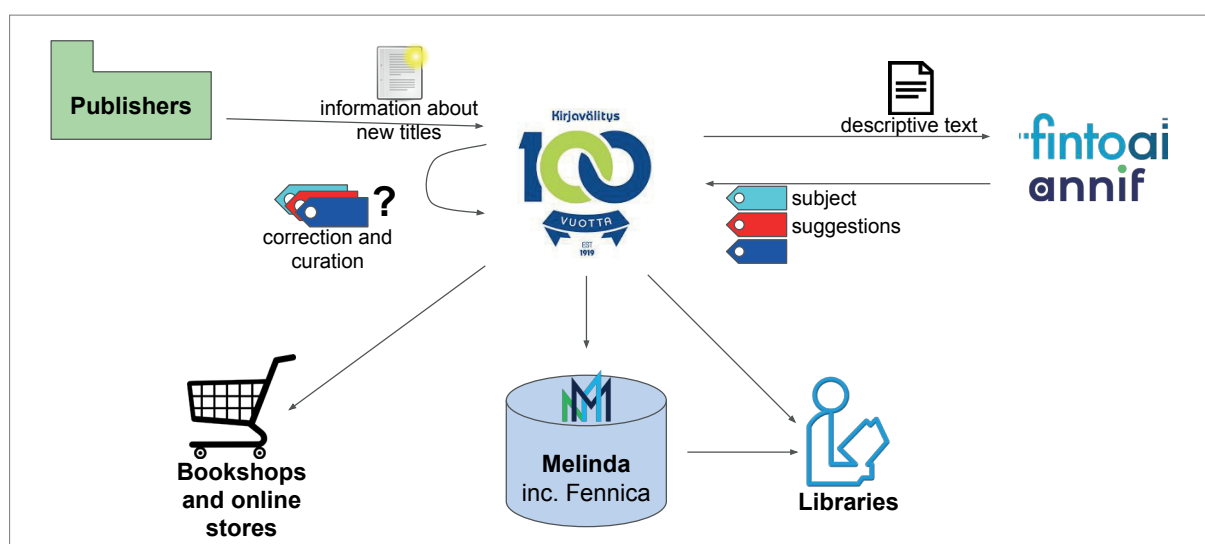
Fig. 7. The book distributor Kirjavälitys Oy receives information from publishers, enhances it with subject indexing assisted by Finto AI, and produces widely used metadata

---

19 https://www.kirjavalitys.fi/en/home/

JLIS.it

Standalone Annif installations

To have more control on the indexing, e.g., for using a custom vocabulary or achieving better indexing quality on a specific topic area, or to support a language not available in Finto AI, a user can install and set up Annif by themselves and train their own models. Training a well-performing Annif model requires possessing adequate amounts of suitable training data, and can be computationally heavy. Searching for good hyperparameters for a model takes a lot of computation time. In contrast, when a model has been trained, and it is used by an Annif instance to offer subject indexing functionality via API, much less CPU resources are needed. For these reasons it can be worthwhile to have separate computing environments for training Annif models and for serving them.

Here we present some institutions that have set up their own Annif installations.

The Leibniz Information Centre for Economics ZBW has a long history of developing automated subject indexing solutions. Currently they are working on the AutoSE project with the aim of transferring their existing automation solutions into productive use (Kasprzik 2020). They use Annif as a part of their framework, and also actively contribute to the development of Annif.

The Finnish Broadcasting Company Yle is setting up Annif for semi-automatic subject indexing of online news articles. They use their own vocabulary and training corpus, and as their custom vocabulary evolves rapidly, they retrain their Annif models every week (Nikkarinen 2021).

The Finnish National Audiovisual Institute (KAVI) offers various services, such as film digitization, and maintains archives. They are also responsible for content rating and screenings of audiovisual material in Finland. KAVI first tested Annif as a standalone installation for indexing radio and TV programs using a speech-to-text transcript of the audio content. Based on the test results, KAVI decided to adopt Annif for this use in their future archive management system (Lehtonen and Piukkula 2020).

The National Library of the Netherlands has explored the possibilities of automatic indexing (Kleppe et al. 2019). Annif is now used as a part of their larger tool that is being developed for library cataloguers (Haighton and Veldhoen 2020). The training data for their current models have been gathered from a collaborative cataloguing system for Dutch libraries. The data consists of titles, subtitles and summaries of Dutch e-books. The Brinkman thesaurus[20] has been used as the controlled vocabulary. Annif has also been applied in a Dutch research project called Entangled Histories. The project focused on early modern ordinances, i.e. law texts, and Annif was used in their classification (Romein, Veldhoen, and de Gruijter 2020).

Dissemin[21] is an online service for researchers to find open publishing repositories for their publications. Dissemin uses Annif to categorize academic pre- and postprints uploaded to open repositories.

---

[20] https://www.kb.nl/sites/default/files/docs/brinkmanonderwerpen-2018.pdf
[21] https://dissem.in/

JLIS.it

Community building

We aim to foster a community around Annif and to make it easy for people to learn about it. Annif has a website that serves as an introduction and an interactive demo. In the Annif GitHub project, we offer a thorough technical description and tips for Annif use. Users can also report bugs or contribute ideas and solutions using GitHub issues and pull requests. There is also a user forum called annif-users[22] where people can ask for help, discuss and share their experiences. The forum is also a platform for Annif-related announcements and news.

Together with ZBW, we have created a hands-on tutorial[23] to help people get started with Annif. The first tutorial session was held at the SWIB19 conference[24]. When the Covid-19 pandemic hit in 2020, we turned the material into an online tutorial suitable for self study, with videos on You-Tube and exercises on GitHub. We have organized several interactive workshops based on the tutorial materials at suitable online conferences.

We also took part in the EU-funded High-Performance Digitisation project, which was a joint effort with CSC – IT Center for Science and the National Archives of Finland. The project sought to find intelligent solutions for automatic indexing workflow in LAM organizations. We were really pleased with this collaboration, which resulted in e.g. the discovery and thorough evaluation of the highly efficient Omikuji algorithms that were later integrated into Annif. The project is described on its web page[25] and in Lehtinen & Kallio (2020). The project also produced a whitepaper (in Finnish) describing the uses and challenges of automatic subject indexing in a cultural heritage organization, with Annif as an example (Hulkkonen et al. 2021).

Conclusion and Lessons Learned

Manually indexing documents for subject-based access is a labour-intensive process, and with the growing mass of digital material it becomes more and more difficult to keep up. There is a need for automation. Although it has taken several years and a lot of development effort, we have successfully created an open source solution for multilingual, vocabulary independent automated subject indexing that has become a production service used in many Finnish libraries, especially through the Finto AI service.

Annif is a unique framework into which different text classification algorithms can be integrated. The algorithms may be used alone, or in combinations called ensembles. We have found that the ensembles nearly always perform better than the individual algorithms.

Subject indexing is not an easy process, either for human indexers or for algorithms. Some parts of it are inherently subjective. When humans do subject indexing, they can have very different perspectives, or sometimes simply make mistakes. These types of mistakes or differences of opinion, however, are usually still relatable or understandable. When algorithms do subject indexing, their mistakes often do not necessarily make any sense from a human perspective.

---

[22] https://groups.google.com/g/annif-users
[23] https://github.com/NatLibFi/Annif-tutorial/
[24] https://swib.org/swib19/
[25] https://www.csc.fi/en/-/high-performance-digitisation

JLIS.it

There are many approaches for evaluating the quality of automated subject indexing systems. We have found that a combination of approaches works well for our purposes. Quantitative comparisons to a human indexed gold standard are the easiest to produce, and we perform them frequently both for the purpose of algorithm development and for evaluating the models that we deploy into production services. User oriented evaluation methods, such as assessment by evaluators, are more laborious, but they produce important insights about how algorithmically produced subject indexing differs from manually created indexing. Organizing workshops around automated subject indexing has provided a way of crowdsourcing the human evaluation effort, while simultaneously spreading awareness about automated indexing among librarians. We have also started to track how our tool is being used in the indexing workflow of systems that are using our API services. In the future, it would also be possible to investigate how the use of automated indexing affects users of retrieval systems.

The Annif tool is increasingly being deployed in Finnish library systems by integration with the API services provided by Finto AI. The Finto AI web user interface is also being used directly by librarians in cases where direct integration between systems is not feasible or has not yet been implemented. So far, users have been very positive towards the subject suggestions given by the service, as it provides an initial suggestion of potential subjects instead of an empty field to fill in. This is especially important for university library repositories where students, who are usually not experts in subject indexing, upload their own thesis documents.

The API services available through Finto AI are currently limited in the terms of indexing vocabularies and languages we can offer. We are working with Finnish organizations that have more diverse needs, for example custom domain-specific vocabularies, so that we can expand the service in the future.

Annif has been community oriented open source software from the start. We have created a web site and a wiki with technical documentation, set up a user forum, presented the tool at conferences and webinars, and together with ZBW, produced a tutorial for learning the basics of the tool. The effort put into community building is starting to pay off, as we are seeing an increasing number of test installations of Annif and some organisations are investing seriously in the adoption of Annif, for example by making extensive tests and comparisons.

One of the challenges in adopting Annif is collecting suitable training data and converting it to the corpus formats that Annif understands. This process usually requires programming skills. Even with a corpus in the correct format, achieving and maintaining good quality can be a challenge. We have gathered advice for setting up and refining projects into a wiki page[26].

There are upsides and downsides of the open source model for library systems. It allows for freedom and flexibility, but requires more technical expertise and resources than similar systems and services provided by commercial vendors. Organizations adopting an open source solution must be prepared to build the in-house expertise required to set up and maintain the systems. Some of the development effort can be shared and pooled through co-operating using code sharing platforms such as GitHub.

---

[26] https://github.com/NatLibFi/Annif/wiki/Achieving-good-results

In the future we continue to actively develop Annif and Finto AI. We hope to keep the community involved and welcome any contributions and feedback. Our aim is to support more vocabularies and languages in the Finto AI service while following the development of new text classification algorithms and utilizing them.

Acknowledgements

JLIS.it

References

Golub, Koraljka, Dagobert Soergel, George Buchanan, Douglas Tudhope, Marianne Lykke, and Debra Hiom. 2016. 'A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval'. *Journal of the Association for Information Science and Technology* 67 (1): 3–16. https://doi.org/10.1002/asi.23600.

Haighton, Thomas, and Sara Veldhoen. 2020. 'Assisted Keyword Assignment Using Annif. KB Lab: The Hague.' 2020. http://kbresearch.nl/annif/.

Hulkkonen, Juha, Juho Inkinen, Aleksi Kallio, Markus Koskela, Mikko Lappalainen, Mona Lehtinen, Mats Sjöberg, Osma Suominen, and Laxmana Yetukuri. 2021. 'Sisällönkuvailun automatisoinnin haasteita ja ratkaisuja kulttuuriperintöorganisaatiossa'. Kansalliskirjaston raportteja ja selvityksiä. http://urn.fi/URN:ISBN:978-951-51-7233-4.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. 'Bag of Tricks for Efficient Text Classification'. *ArXiv:1607.01759 [Cs]*, August. http://arxiv.org/abs/1607.01759.

Kasprzik, Anna. 2020. 'Putting Research-Based Machine Learning Solutions for Subject Indexing into Practice'. In *Proceedings of the Conference on Digital Curation Technologies (Qurator 2020)*. Berlin, Germany. http://ceur-ws.org/Vol-2535/paper_1.pdf.

Khandagale, Sujay, Han Xiao, and Rohit Babbar. 2020. 'Bonsai: Diverse and Shallow Trees for Extreme Multi-Label Classification'. *Machine Learning* 109 (11): 2099–2119. https://doi.org/10.1007/s10994-020-05888-2.

Kleppe, Martijn, Sara Veldhoen, Meta van der Waal-Gentenaar, Brigitte den Oudsten, and Dorien Haagsma. 2019. 'Exploration possibilities Automated Generation of Metadata'. Zenodo. https://doi.org/10.5281/zenodo.3375192.

Lehtinen, Mona, Juho Inkinen, and Osma Suominen. 2019. 'Aaveita koneessa: Automaattisen sisällönkuvailun arviointia Kirjastoverkkopäivillä 2019'. *Tietolinja* (blog). 2019. http://urn.fi/URN:NBN:fi-fe2019120445612.

Lehtonen, Tommi, and Juha Piukkula. 2020. 'Automaattinen asiasanoitus Radio- ja televisio-ohjelmatietokanta Ritvassa'. *Informaatiotutkimus* 39 (1): 27–45–27–45. https://doi.org/10.23978/inf.88107.

Medelyan, Olena. 2009. 'Human-Competitive Automatic Topic Indexing'. Thesis, The University of Waikato. https://researchcommons.waikato.ac.nz/handle/10289/3513.

Niininen, Satu, Susanna Nykyri, and Osma Suominen. 2017. 'The Future of Metadata: Open, Linked, and Multilingual – the YSO Case'. *Journal of Documentation* 73 (3): 451–65. https://doi.org/10.1108/JD-06-2016-0084.

Nikkarinen, Irene. 2021. 'Annif <3 Yle 2.0: Annifin osittainen käyttöönotto artikkeleiden koneavusteisessa asiasanoituksessa'. Presented at the Meeting of the Finnish Automatic Indexing Interest Group, March 15. https://www.kiwi.fi/display/tekoalykumppanuus/Automaattisen+kuvailun+verkoston+tapaamiset?preview=/147358597/211911484/Automaattisen%20kuvailun%20verkoston%20tapaaminen%2015.3.2021%20Annif.pdf.

JLIS.it

Prabhu, Yashoteja, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. 'Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising'. In *Proceedings of the 2018 World Wide Web Conference*, 993–1002. WWW '18. Lyon, France. https://doi.org/10.1145/3178876.3185998.

Romein, C. Annemieke, Sara Veldhoen, and Michel de Gruijter. 2020. 'The Datafication of Early Modern Ordinances'. *DH Benelux Journal* 2. https://journal.dhbenelux.org/journal/issues/002/article-23-romein/article-23-romein.html.

Stevens, Mary Elizabeth. 1965. *Automatic Indexing: A State-of-the-Art Report*. NBS Monograph 91. Washington, D.C: United States. Government Printing Office.

Suominen, Osma. 2019. 'Annif: DIY Automated Subject Indexing Using Multiple Algorithms'. *LIBER Quarterly* 29 (1): 1. https://doi.org/10.18352/lq.10285.

Suominen, Osma, and Pia Virtanen. 2020. 'Yle Meets ANNIF – an Open Source Tool for Automated Subject Indexing'. Presented at the EBU MDN Workshop 2020, June 10. https://tech.ebu.ch/contents/publications/events/presentations/mdn2020/yle-meets-annif--an-open-source-tool-for-automated-subject-indexing.

Toepfer, Martin, and Christin Seifert. 2020. 'Fusion Architectures for Automatic Subject Indexing under Concept Drift: Analysis and Empirical Results on Short Texts'. *International Journal on Digital Libraries* 21 (2): 169–89. https://doi.org/10.1007/s00799-018-0240-3.

Uhlmann, Sandro. 2020. 'Automatische Vergabe von GND-Schlagwörtern Mit Annif - Ergebnisse Einer Evaluation Im DNB - Projekt EMa'. Presented at the Erfahrungen und Perspektiven mit dem Toolkit Annif, December 3. https://wiki.dnb.de/display/FNMVE/Workshop+2020%3A+Toolkit+Annif.

Wilbur, W. John, and Won Kim. 2014. 'Stochastic Gradient Descent and the Prediction of MeSH for PubMed Records'. *AMIA Annual Symposium Proceedings* 2014 (November): 1198–1207.

JLIS.it

Towards an open and collaborative Authority Control

Barbara Katharina Fischer[(a)]
with the cooperation of Jürgen Kett[(b)]**,**
Sarah Hartmann[(c)]**, Mathias Manecke**[(d)]

a) Deutsche Nationalbibliothek (The German National Library)
b) Deutsche Nationalbibliothek (The German National Library)
c) Deutsche Nationalbibliothek (The German National Library)
d) Deutsche Nationalbibliothek (The German National Library)

ABSTRACT

As digital transformation is speeding up, the need for a reliable retrieval is too. Libraries have long used *authority files* to enhance the search for information. Now, as the entire GLAM field is increasingly presenting its content online, national libraries face the requirement to provide authority data as reference points to a far more diverse community. The request is not limited to persistent identifiers but new records on non-librarian entities are needed. The German National Library (DNB) aims to provide an open framework that allows *collaboration* on all levels: editing the records, defining the regulations and standards plus ease the data flow in both directions. To this end, the DNB has started an ambitious project transferring both the authority file records and their regulations into a *Wikibase* instance. The article relates the findings working with the beta version of the software that drives *Wikidata*. To spur the process the DNB co-published the WikiLibrary Manifesto together with Wikimedia Deutschland. The institutions signing the manifesto shall cooperate to improve the building of a technical infrastructure that will ease knowledge equity through the *FAIR Data Principles* and the creation of a structured data ecosystem. The manifesto was signed by IFLA in June 2021.

JLIS.it

> "What really distinguishes us is the way in which we collaborate on a major scale."[1]

When people discuss topics with verve and persistence, this is generally a sign of dedication and connection. The topic of "Opening the GND" features these positive qualities. It affects and moves many people. It raises questions on the major topic of collaboration, both in great detail and a vast range of different contexts. The opening quote to this article is taken from the historian Yuval Noah Harari's[2] much-acclaimed graphic novel "Sapiens", which tells the history of how humankind developed. It also describes our work in the *Office for Library Standards* (AfS) at the German National Library. Organising collaborations is at the heart of what we do. Our task is to facilitate the cataloguing of knowledge resources across national and disciplinary boundaries. We organise collaborations by promoting consensus on standards that we ultimately use to describe the world while keeping them equally comprehensible for all. Using these standards, the community of German-language libraries is defining how publications should be described with greater precision than by means of natural language so that others can definitively refer to them. This is where the *Integrated Authority File* (GND) comes into play. Harari refers to the nature of humankind as a whole, and how this differs from the character of chimpanzees, for example. The work of cataloguing, the definition of media based on the rules of descriptive and content cataloguing, is far removed from the challenges faced by Homo Sapiens during the Stone Age. And yet, in a sense, it is simply a different section of the same light beam. As a result of the "cognitive revolution"[3] that occurred back then, today, we are facing the challenges of the digital transformation. And this too we will master precisely because of our ability to collaborate. This is what we do.

In the course of opening the GND to include communities beyond library institutions, one thing has become ever clearer: the GND is much more than just a collection of nine million authority data records on people, places, corporations, conferences, works and subject headings.[4] It also describes an organisational structure that reflects the state of its current users. It refers back to a certain data model that is based around the needs of its users. It is subject to specific rules and can be regarded as a specialist tool within an specialised environment defined by the requirements of the library community. Yet the new user groups are organised differently. They have other data models. They catalogue the objects of their interest according to different rules and use a different technical infrastructure. And yet they are still very interested in using the GND authority data. They don't just wish to use the identifiers in their cataloguing work, but also want to be able to create new GND data records when they see a need to do so. They want to become an active part of the GND community. To this end, we need to work together to consider carefully what we can change, and how much, without damaging the core of the GND. This is because everyone wants to preserve its reliable quality. Our task is once more to organise our collaborative efforts in line with a collective intentionality.

---

[1] Quote taken from Harari 2020, p. 68.

[2] Harari 2020.

[3] On the concept of the "cognitive revolution", cf. Harari 2012, pp. 11-100.

[4] The record type *Conferences* in the GND makes particularly apparent how interwoven the GND is with its users in the world of libraries. That is because this record type describes a specific kind of publisher. See all categories in the GND ontology: https://d-nb.info/standards/elementset/gnd

JLIS.it

An instrument for broadening participation

Opening the GND is like the concert given by an entire orchestra of stakeholders and activities. One instrument in this orchestra, a starting point for a careful adaptation, is the technical environment in which the GND is rooted. It is not the notion of dispensing with the existing technical infrastructure, but much more the idea of offering a parallel infrastructure, that has drawn our attention to the database software Wikibase[5]. Wikibase is a piece of open-source software from the Wikimedia Foundation. This foundation has previously developed the Mediawiki software, which is used to operate millions of Wikis around the world. The most famous Wiki is Wikipedia, operated by Wikimedia. The Wikidata project was launched nine years ago with the aim of improving Wikipedia. A database for structured data with which one can describe the world in a way that can be read by both humans and machines alike. The software empowering Wikidata is Wikibase. Wikibase features certain properties designed to make large-scale collaboration easier:

- It offers web-based access.
- It facilitates parallel collaborative working.
- It automatically logs the version history and its editors.
- It offers a dedicated discussion page for every data record.
- It is geared towards multilingual user communities.
- It offers a simple and flexible (though also limited) data model.
- Entering new content works easily and intuitively.

We intensively studied these properties at the German National Library in 2019 and summarised our conclusions in our evaluation[6] in collaboration with Wikimedia Deutschland. In this context, we also explored current weaknesses in the system and potential areas for development. In its current iteration, the system falls far short of meeting all the requirements for an ideal editing system and hub for cultural institutions. To this end, it still needs to outgrow its origins as a piece of Wikidata software. Nevertheless, we were able to identify the fundamental prerequisites for its productive use in the context of the AfS. What matters is less the current status of the product and more the inherent potential in its further development and the establishment of a broad community in the cultural sector.

In 2020, we first considered how to make the most effective use of Wikibase in broadening participation in the GND, before creating the conditions for implementing our plans as efficiently as possible. We decided to become active on three levels. We want to:

- Create a second home for the GND as an authority file within a Wikibase database. New user communities can make suggestions for new GND data records more easily and independently of the existing technical structures, and compare their data to the GND with greater ease in order to avoid duplication.
- Create partnerships with Wikimedia and other institutions also wanting to use Wikibase, in order to collaborate on improving the software so as to ultimately establish an ecosystem for cultural data and research data.
- Thirdly, we want to re-order the very frameworks underpinning the GND and our cataloguing work, make these more accessible and easier to adapt to any changes.

---

[5] Link to the Wikibase website: https://wikiba.se/

[6] Link to the blog post on our evaluation: https://wiki.dnb.de/pages/viewpage.action?pageId=167019461

JLIS.it

The second home of the GND

In the world of libraries, the GND has long served as a referencing and rationalisation tool, as did the four authority files that preceded it. It is integrated into certain frameworks and proprietary software structures that are, however, relatively inaccessible to users from outside the world of libraries. We believe that we can use Wikibase to make it easier for some of these target groups to collaborate on the GND.

To this end, we wish to import all the existing GND data records and their corresponding links into a Wikibase entity in 2021. This may sound like a simple task. However, Wikibase's importation interfaces are still very much aligned with the needs of Wikidata. For this reason, we have sought professional support from a Wikibase specialist, who is assisting us as a service provider in the transfer of the database infrastructure, the data importation and the creation of user-friendly input screens. In the next step, we will then invite experienced and new GND users to test the data-entry and search processes in the new environment so that we can further improve these.

During the second half of 2021, we are planning a technical workflow for synchronising the GND Wikibase entity with the CBS system[7]. The plan is to enable new and existing users without any WinIBW[8] access to enter their data as a suggestion in the Wikibase entity.

One long-term goal is to offer a user-friendly and supportive data-recording environment for the GND. The *GND web forms*[9] represent a first step in this direction, as they are considerably more user-friendly than the data-entry systems used by libraries. The web forms currently can be used to record people and corporations. However, this currently envisaged approach is not flexible enough. In addition to the two aforementioned GND record types, there are four more. These six record types unite approximately 50 entity codes[10], each with specific properties via which the respective entities can be recorded as GND data records. These would require a dynamic entry form that adapts to the entity type or usage context chosen, offers necessary and typical entry elements, highlights useful entries and thus guides the user through the entry process. It remains to be seen whether Wikibase represents the right platform for this in the medium term. At present, Wikibase lacks such features. For now, no update to the generic entry interface is planned "ex works". There is also no option of limiting the offering to fundamental elements or values. The user is always confronted with the full range of properties and values, and isn't offered any assistance in decision-making. One aim for 2021 is to establish whether this can be facilitated via the development of a Wikibase expansion, and also which changes would have to be implemented in Wikibase by Wikimedia in order to more adequately support the creation of customisable entry forms that assist the user.

---

[7] CBS: proprietary library data-entry software from OCLC.

[8] WinIBW: licensed software for entering data in the GND.

[9] The GND web form for persons and corporate bodies is specifically intended for users from cultural institutions such as smaller libraries, archives and museums who would like to create or modify small quantities of data records in the GND. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_Webformular/gnd_webformular.html

[10] Details on entity coding in the GND https://wiki.dnb.de/download/attachments/90411323/entitaetenCodes.pdf

JLIS.it

The WikiLibrary Manifesto

Another area our work will focus on in 2021 is our partnership with Wikimedia Deutschland and other institutions in order to improve Wikibase as a technical infrastructure. Adherence to the FAIR Data Principles (Findability, Accessibility, Interoperability and Reusability)[11] when providing data is becoming increasingly important in an ever-growing number of contexts. Data are to become more interlinked in order to make it easier overall to generate new knowledge. This especially applies to data that were generated using public funding. This represents a great challenge for many institutions. It raises the question as to whether they should offer their data in a collective pool for structured data, like data portals. Such institutions must ask themselves whether they are willing to face all the potential consequences, such as sacrificing control over the data model, data-recording rules and quality-assurance processes. Or should they instead use stand-alone solutions and thus accept that their data will be less visible and get re-used less? By broadening participation in the GND, we wish to create alternatives. We are committed to creating a reliable, machine-readable and communally managed Linked Open Data Network for the arts, sciences and culture as a viable basis for FAIR knowledge. Instead of a central platform, we favour an open network of interlinked databases. This requires a communal organisational framework. We wish to provide this within a single network. A network is only ever as good as the partners within it. To this end, the German National Library co-published the WikiLibrary Manifesto together with Wikimedia Germany. Allmost forty institutions have already accepted our invitation. The manifesto invites the undersigning institutions to collaborate on the basis of the following principles:

- Promoting free licenses for data and their software environment.
- Shaping spaces where diverse communities thrive. (Community gardening).
- Providing structured data based on FAIR data principles in order to be able to transparently transform data into information to create FAIR knowledge.
- Promoting common core standards created consensually and collaboratively.
- Providing open governance structures and embedding them into existing systems.
- Dedicating resources to obtain user interfaces that are accessible to and user-friendly for everybody who wants to contribute and actively care for data and knowledge.
- Fostering data literacy in the digital transformation on the three stages: data, information and knowledge.

Of equal importance, if not more so, is the communal implementation of specific measures by all signatories in partnership with Wikimedia Germany. The aim is to promote Wikibase as a promising technical infrastructure for the storage, editing and exchange of data on the basis of the FAIR Data Principles. We wish to shape Wikibase into a user-friendly reference-database software for data hubs in order to promote the desired data ecosystem. To this end, we are inviting further institutions from the world of libraries, from all GLAM (galleries, libraries, archives and museums) areas and from the humanities to use Wikibase in order to create an ecosystem of structured data that comes closer to a true semantic web for FAIR knowledge.[12]

---

[11] Information on the Fair Data Principles https://www.forschungsdaten.org/index.php/FAIR_data_principles

[12] As an institution, you can co-sign the manifesto via a simple form by following this link: https://www.wikimedia.de/projects/wikilibrary-manifest/

JLIS.it

The DACH documentation platform[13]

Would it have occurred to you that the article opposite about a football match was written by a computer program? In recent years, the results of computer linguistics have evolved ever further with the aid of artificial intelligence. Writing programs draw content from structured databases and construct the texts with the individual components according to certain specifications. This is the backdrop to our deliberations on recording the frameworks for descriptive and content cataloguing[14] and the data-entry guidelines for the GND as structured data within a Wikibase entity. For decades now, we have been issuing extensive, detailed texts with precise instructions on which data fields must be entered in the GND, for example. Underpinning these texts are the frameworks for descriptive and content cataloguing, the requirements and limitations of the respective software used for cataloguing, and ultimately also the requirements for the exchange of data. Each time a detail is amended at any point in this complex network, the same amendment must also be implemented in many texts that refer to said point. In each instance, this requires labour- and time-intensive research in a large number of PDF pages. Another consequence of this form of knowledge management is that lots of detailed information – such as how to enter a date, for example, or how to record a job description, or which code to use for which country – has to be repeated in various places to avoid having to hunt for said information. When making an amendment, it is important to maintain an overview of every other area impacted by that amendment. There is an inherent risk of errors and a lack of transparency. It certainly makes any guidance less user-friendly, as there is a continuous need for amendments.

The basic principle is strikingly simple. Let us first focus on the GND itself. The number of fields with which one can describe entities for authority data records in the Pica or Marc 21 data formats is manageable at around 300. These data fields or elements serve to make statements regarding

---

[13] DACH documentation platform: The platform is designed to bring together all the frameworks for library-based cataloguing and the data-entry guidelines for the GND in the German-speaking regions (Germany, Austria and Switzerland).

[14] This refers to the RDA and RSWK frameworks

JLIS.it

the properties, relationship types, sub-categories or entity codes for the respective entities being described. The data elements contain defined characteristics and different codes, depending on the data format. If all the elements are stored in a corresponding database, the data elements can be assembled in modular fashion, just like a construction kit, according to the rules of the frameworks the database is organised around.

The data-entry guidelines for people alone, with all the requisite entity codes in the GND, encompass 46 pages.[15] Yet the elements to be recorded are few. Along with the name, the other primary elements are the date of birth and death, and any links to other databases, such as place names in the form of the place of birth, the place(s) where the individual worked, or similar. For each of the entity codes in the record-type "persons, a new description is provided each time of how the element "place" must be modelled, for example. If these definitions were to be stored in a database, as a rule, one could simply enter the respective element. This means that if the rule governing the characteristics for recording a regional corporation changes,[16] one can change this centrally in a single location, and all other locations where this element is used are automatically updated too. It is the same principle as applied in the authority data records of library catalogues.

We have started to describe in a structured format all the elements used in the GND. To do so, we are adopting the specifications contained in the frameworks. Now the challenge will consist in writing comprehensible continuous texts in which one can sensibly embed the elements. These can then be updated more concisely than before, and also potentially serve as the foundation for the creation of entry forms for the database with all GND data records.

Sometimes it is beneficial to reflect on the sense and purpose of one's work in order to remain motivated, stay focused and convey to others why this work is important and deserves funding. With this workshop report, we wish to provide you with an insight into our work and the ideas behind it. An exciting time of pioneering work lies ahead of us. This work is made even more interesting thanks to the other, concurrent Wikibase projects in the newly formed consortia of the National Research Data Infrastructure Initiative (NFDI) and other major universal and national libraries in Europe and America, with whom we are in close contact. We will keep you up to date on the latest developments.

---

[15] Also cf. https://wiki.dnb.de/pages/viewpage.action?pageId=90411361&preview=/90411361/94831186/EH-P-01.pdf

[16] A local corporation is an entity code from the group of geographic entities or places.

JLIS.it

References

Harari, Yuval Noah. 2013. *A brief history of humankind.* Munich: Dt. Verlags-Anstalt.

Harari, Yuval Noah. 2020. *Sapiens. The birth of humankind. Graphic novel.* Munich: C.H. Beck.

JLIS.it

Wikidata: a new perspective
towards universal bibliographic control*

Carlo Bianchini[(a)], Lucia Sardo[(b)]

a) Università degli studi di Pavia, http://orcid.org/0000-0002-6635-6371
b) Università di Bologna. Campus di Ravenna, http://orcid.org/0000-0001-6480-759X

———————

Contact: Carlo Bianchini, carlo.bianchini@unipv.it; Lucia Sardo, lucia.sardo@unibo.it
Received: 10 April 2021; **Accepted:** 22 May 2021; **First Published:** 15 January 2022

———————

ABSTRACT

Traditional UBC provides for the standardization of bibliographic records, the creation of guidelines dedicated to national bibliographic agencies, the creation of the UNIMARC format, and the curation of authority data. Bibliographic Control has deeply evolved since IFLA theorization during the Seventies of the XX Century, due to the availability of a very large range of new bibliographic tools. At the beginning of the XXI century, UBC is quite different and involves new actors. Among these, Wikidata has a background greatly different from that of libraries as institutions: it is not devoted to bibliographic data, nor it is limited to personal authority control, but its value in AC tools like VIAF and National Libraries authority files is undiscussed. After a presentation on how Wikidata items describe and identify bibliographic entities, the authors underline how the existence, use and reuse of Wikidata affect the way the professional community thinks about UBC. Wikidata is a clear example of the need for a new approach to identification and description, that are deeply intertwined. Secondly, from a Wikidata perspective, the relevance of globally preferred and variant access points is lessened. Moreover, descriptions in Wikidata – although conceptually very similar to the traditional one – present differences and potentialities that a traditional description does not have and cannot have. Also from a theoretical perspective, Wikidata offers a pragmatic way to think globally and act locally. In fact, it shows that there is no need for one standardization of practices for establishing the headings and structure of authority records in one international form; instead, users' convenience can be achieved by a technological infrastructure capable to present to each user the information about an entity in its own language and script. Additionally, Wikidata is the most evident example of the distributed and diffused approach of the semantic web to the issue of the universal identification of the entities. Also, Wikidata identification and description show that authority and bibliographic control must be tackled as just a part of the more general topic of the creation of a knowledge graph of all human knowledge by means of linked open data. Lastly, this objective cannot be achieved only by contribution, cooperation, and networking of large national agencies (as in VIAF), as a larger number of stakeholders must be involved to achieve a UBC also including the full indexing of any kind of scientific communication

———————

* The authors cooperated in the redaction and revision of the article. Nevertheless, each author mainly authored specific parts of the article: Carlo Bianchini: sections 1.2, 2, 4, and 5; Lucia Sardo: sections 1, 3, 4.2, and 5.

———————

JLIS.it

1. Introduction: From UBC to the semantic web

The idea of bibliographic control, that is, the idea of being able to have an exhaustive overview of what is published ("the mastery over written and published records which is provided by and for the purposes of bibliography"; Unesco and LC Bibliographical Survey 1950, 1), at least as old as bibliography, took on new connotations in the 1970s, with the birth of IFLA's Universal Bibliographic Control program.

The main objective of universal bibliographic control is the availability of bibliographic records of publications produced in all countries. In the context of the program for the UBC, the emphasis is placed not only on the universality of such control, but also on the standardization of the content of bibliographic records, on the need to have a specific program dedicated to this objective to foster cooperation at the international level and to achieve the goal of a worldwide record of what is published, and on the importance given to the rapid availability of these data.

The program for the UBC has its basis in standardization and in the direct involvement of national bibliographic agencies, a fundamental element for international cooperation: during the 1977 congress devoted to national bibliographies, their tasks were defined to be the documentation of national editorial production, and the drafting of authority records for national authors.

The principles on which universal bibliographic control is based, which have been progressively brought into focus through studies and initiatives, can be summarized as follows: first, the aims of the system are the control and exchange of bibliographic information; worldwide coverage is guaranteed by the cooperation of the national components of the system; the main objective is to make the bibliographic data of all publications universally and promptly available, in an internationally accepted standard format. To achieve this goal, the complete bibliographic record of each publication should be made once only in the country of origin by a national bibliographic agency, in accordance with international standards that permit exchange. In this perspective, the national bibliographic agency, usually established in national libraries, which usually benefit from the mandatory deposit of printed matter, results to be the most appropriate structure for authoritatively identifying and recording the authors and publications of each country, and is responsible for producing a current national bibliography, in which to publish such records as soon as possible, and for distributing such records in various standard formats. The agencies then come to be integrated into an international system and regularly exchange records made.

For UBC principles to be implemented and scaled up, some requirements must be satisfied:

- a canon of principles, standards, and practices governing the creation and structure of catalographic data that is shared on a broad scale must be available.
- each national bibliographic agency must fulfil its responsibilities in a manner that is inclusive of and consistent with accepted standards.
- an infrastructure is needed to support the efficient exchange of data among national bibliographic agencies.

While IFLA's work on satisfactorily scaling up the UBC concept regarding bibliographic records has been successful, regarding authorities it has been largely driven by the recognition of the need to deal with these three critical factors:

- the standardization of practices for establishing the headings and structure of authorities.

JLIS.it

- the promotion of national responsibilities for the creation and "dissemination" of authority records.
- the planning of an infrastructure that supports the effective international exchange of authority records.

The results of the program's work for the UBC are there for all to see: the publication of ISBD, the creation of the UNIMARC format, the publication of authority lists and tools for controlling the forms of personal and collective names, and the Guidelines for the National Bibliographic Agencies and the National Bibliography. Basically, all the work concerning the standardization of bibliographic descriptions and authority records had its basis in the concept of Universal Bibliographic Control.

1.1 UBC for bibliographic and authority data

From the point of view of the UBC, the ISBD standard was first developed, which had the double function of establishing which data were relevant for bibliographic description and in which order they should be presented. This was the first time that such a standardization effort was undertaken; it preceded the formalization of the program but provided for it as a *conditio sine qua non* for its dissemination.[1]

In the same period MARC, a machine-readable format, was created for the exchange of cataloguing information. IFLA, too, considered it essential to develop an international MARC format capable of supporting the exchange of bibliographic data, which is why the development of the UNIMARC format was undertaken, both for bibliographic and authority data. All of this was born, we recall, in a context where catalogs were paper-based, and national needs trumped those of internationalization, especially about the choice of name form for access points.

As Gorman summarizes, "In sum, arriving at a standard set of elements in a standard order and delimited in a standard manner was in the mutual interest of the effort to achieve an international standard for bibliographic description (what became the ISBD); MARC; [sic] and the use of both, each in accord with the other, in achieving national and international standardization, cooperation, and sharing; leading, ultimately, to Universal Bibliographic Control" (Gorman 2014, 826-827).

1.2 Wikidata: a tool of bibliographic interest in the semantic web

In 2011, the Library Linked Data Incubator Group, a working group with the aim "to help increase global interoperability of library data on the Web",[2] published its final report. It was focused on what libraries can do for the semantic web and what the semantic web can do for libraries, and it underlined that libraries had created and curated a relevant amount of rich data that can "help reduce redundancy of bibliographic descriptions on the Web by clearly identifying key entities that are shared across Linked Data" (W3C Incubator Group 2011). The report offered a new perspective on thinking about the relevance, scope, and purpose of Universal

---

[1] For an overview about the origins of ISBD, see (Anderson 1974; Gorman 2014).

[2] https://www.w3.org/2005/Incubator/lld/

Bibliographic Control (UBC), beside to "make universally and promptly available, in a form which is internationally acceptable, basic bibliographic data on all publications in all countries" (Anderson 1974, 11).

Since the publication of the Report, many tools with a top-down approach have been developed for the identification of entities (people, locations, works, and expressions) such as VIAF, ISNI or ORCID. The top-down approach of these tools reflects the role assigned to the national agencies by UBC. Nevertheless, some authors suggest that, in the semantic web, "building collaborative authority registries linked to standardised identifiers is one of the fundamental cornerstones of the new Universal Bibliographic Control" (Illien and Bourdon 2014, 15) and that "a better mix of bottom-up and top-down methodologies" is needed to support all those who wish to think globally and act locally (Dunsire and Willer 2014, 11).

Since 2012, Wikidata has developed as a new global actor of the semantic web with both a bottom-up and very inclusive approach. Wikidata is a freely available hosted platform that anyone – including libraries – can use to create, publish, and reuse LOD (Allison-Cassin and Scott 2018). Its main goal and function are to work as a central storage for many Wikimedia projects, but it is also used in external services, for example in VIAF or in the Google Knowledge graph (Vrandečić and Krötzsch 2014), for the enrichment of the quality of bibliographic records (Nguyen, Dinneen, and Luczak-Roesch 2020), and for bibliometrics projects and tools (Lemus-Rojas and Odell 2018; Nielsen, Mietchen, and Willighagen 2017; Hernández-Cazorla, Ramírez-Sánchez, and Rodríguez-Herrera 2019; Seidlmayer et al. 2020; Mietchen and Rasberry 2020). Moreover, in the last years, the Wikidata role as an important tool for identifying entities has been increasingly reconsidered (Association of Research Libraries 2019, 27; van Veen 2019; Linked Data for Production 2020).[3]

2. Identification in Wikidata

As Wikidata is a central storage for all Wikimedia projects, it aims to record data about any kind of item (i.e., entity) and property relevant for all its projects. For example, items of Wikidata can be geographical places, administrative units, events, architectonic objects, any entity of interest for the user, and, of course, any 'res' provided for by IFLA LRM model.

In fact, Wikidata shows a relevant interest for the bibliographic universe. Statistics show that Wikidata records about 91 million of items, 31,5% (ca 22,5 million) of which are scholar articles, and nearly 9% (ca 6.376.000) of the existing items are of human type (Q5). Anyway, this class includes any kind of humans, such as kings, politicians, football players and so on, and not just authors of literary or scientific works. Nevertheless, items representing an authority record (not just of humans) can be estimated to be around 6,3 million.[4]

For identification purposes, Wikidata assigns to each item both an URI – for example, https://www.wikidata.org/wiki/Q12418 – and a label, a description, and one or more aliases (figure 1).

---

[3] https://www.wikidata.org/wiki/Wikidata:Wikidata_for_authority_control
[4] Personal items with a VIAF identifier are about 2,3 million, but the number of personal items containing at least one identifier of any VIAF source (such as ISNI, LC, GND etc.) are about 6,3 million.

JLIS.it
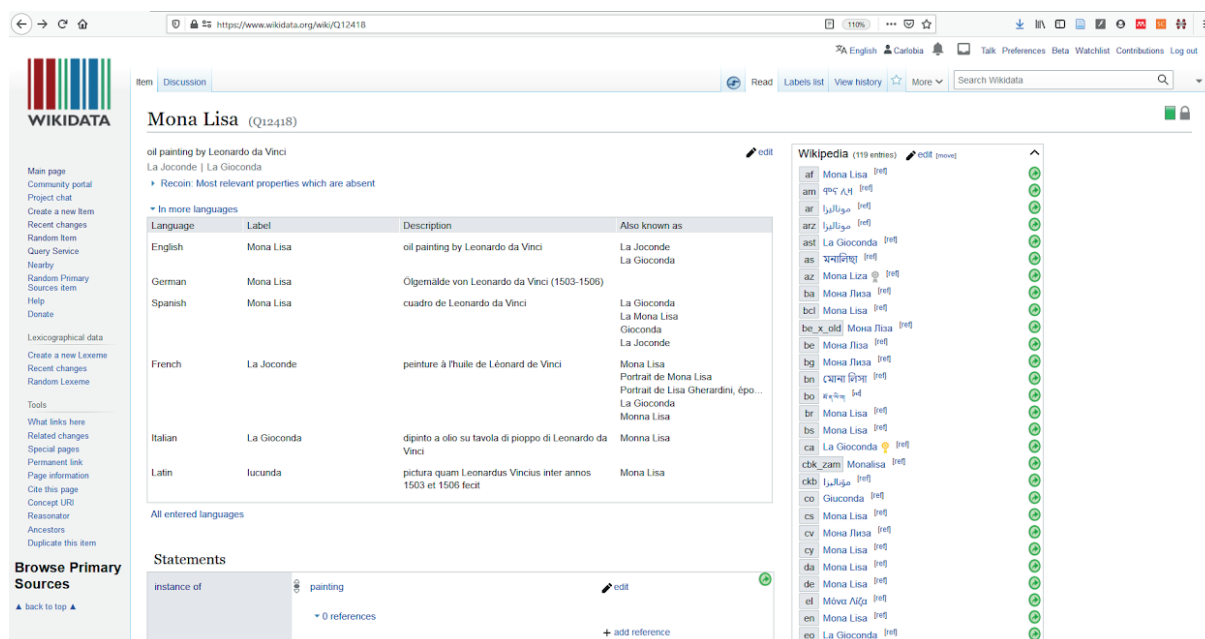
Fig. 1. Example of the main identifying parts (label, description, aliases) of a Wikidata item

The label is the first data element, and it can be considered as the preferred form of the name for the represented entity. In fact, it can be expressed in any existing language and any registered user is enabled to visualize the label in his/her own language and script, if available. Moreover, the preferred form of the name is expressed directly by the user that usually creates the item. So, preferred forms in Wikidata are *literally* founded on common usage and on the convenience of the user provided for by ICP (IFLA Cataloguing Section and IFLA Meeting of Experts on an International Cataloguing Code 2016), and not on these principles *interpreted by a code* of national or international rules! It must be noted that multiple languages and scripts are available for the very same entity, and not for a cluster of nationally created forms like in VIAF.

The second element is the description of the item, that is a short phrase to describe the item. It is in free language and it is useful to quickly distinguish an item from any other item with the same label (for example, "Love"; figure 2), i.e., to disambiguate homonyms.

JLIS.it
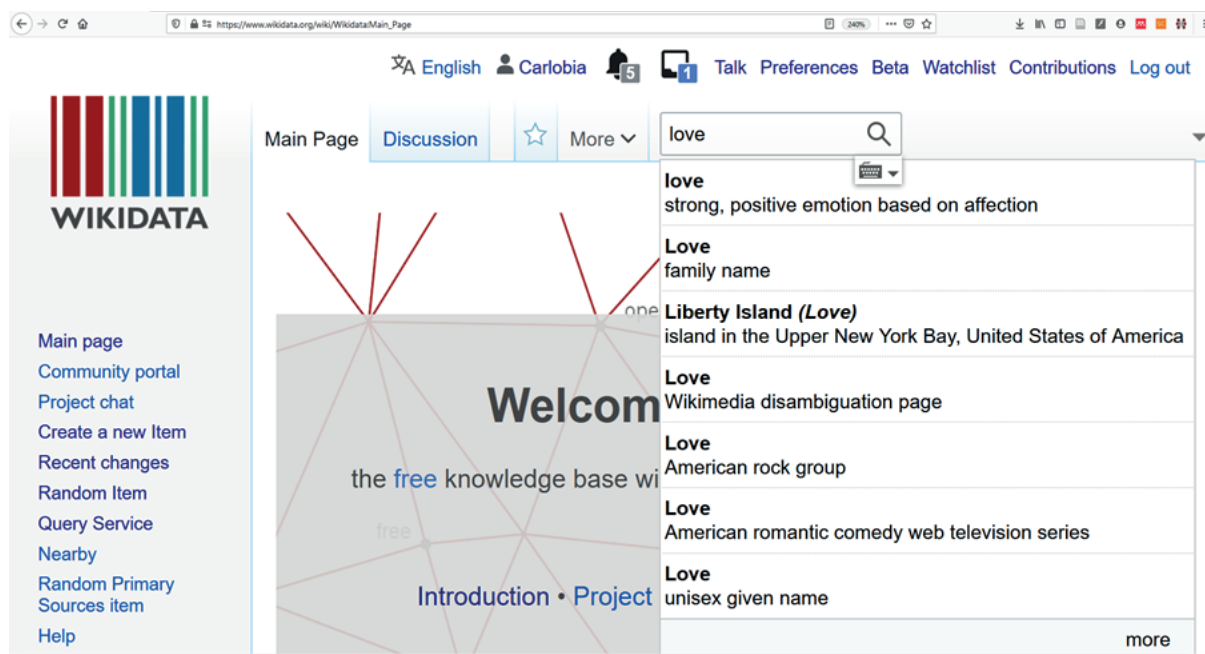
Fig. 2. Descriptions helps to disambiguate items with the same label "Love"

The third element for the quick item identification in Wikidata are the aliases, that are variant forms of the name in one specific language and script (as variant forms of the name in any other language are provided in the form of preferred and variant names in those languages; figure 3).

Fig. 3. Aliases available in multiple languages and scripts for S.R. Ranganathan

JLIS.it

While the unique identification of the entity is based on a neutral URI (for example: https://www.wikidata.org/wiki/Q1334284), both labels and aliases – in any available language and script – work as access points. This pragmatical approach overcomes the theorical ICP and RDA distinction between preferred and variant access points.

All the remaining properties are registered after the identification elements described above. Nevertheless, they are logically divided into main parts: properties and identifiers. While *properties* are traditionally associated with the *descriptive* goal of the data (see below § no. 3), all the other external identifiers respond to the need of the fourth linked data principle stated by Tim Berners-Lee: "Include links to other URIs so that [users] can discover more things" (Berners-Lee 2006).

External identifiers have the goal to interlink the URI of the item of Wikidata with any other identifiable entity described in the semantic web.[5] For this reason, Wikidata is more and more recognised for its relevance in the identification of semantic web entities (Association of Research Libraries, 2020; Linked Data for Production, 2020; van Veen 2020). Enriching data with Wikidata ids allows to discover other sources of data and information available in the semantic web.

To create a link towards an external identifier, a specific property must be created in Wikidata to define that identifier. So, it is possible to know how many identifiers are available for different kinds of entities. In figure 4, created by Simon Cobb (Cobb 2019, 5), the number of identifiers associated with each kind of entity are shown.
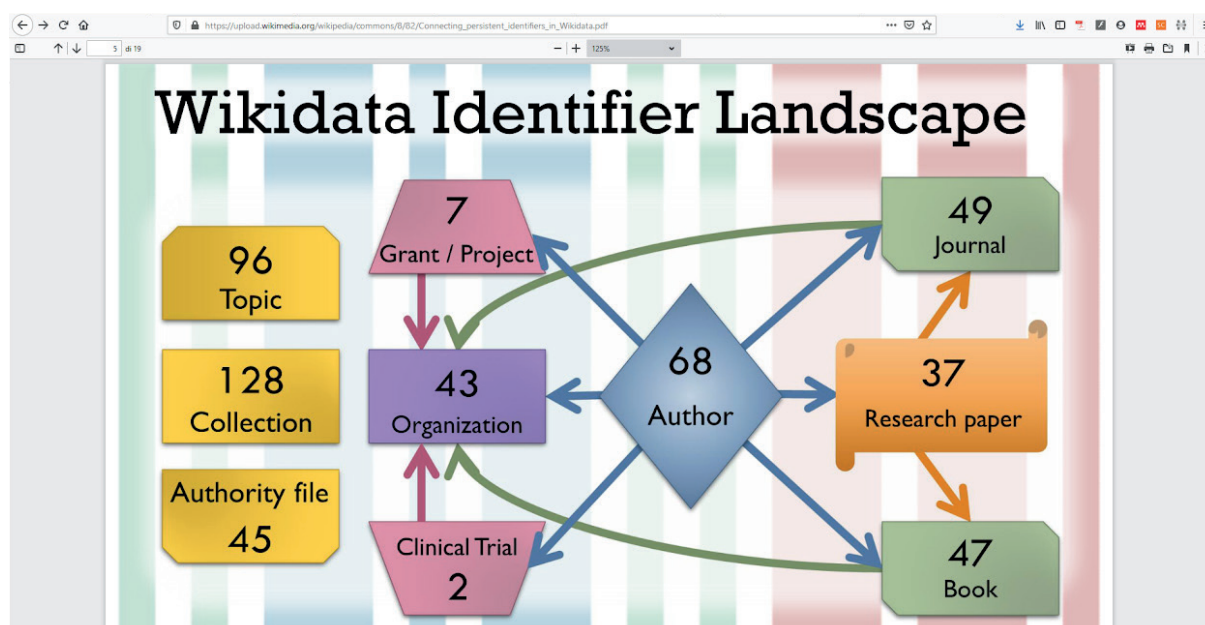
Fig. 4. Number of identifiers in Wikidata for each kind of entity (by Cobb 2019)

---

[5] Entity Explosion is a very interesting tool to understand the potential uses of this WD function for the navigation in our discovery tools; see https://chrome.google.com/webstore/detail/entity-explosion/bbcffeclligkmfiocanodamdjclgejcn.

JLIS.it

Most frequently used identifiers in Wikidata are: PubMedID (60,152,490), DOI (26,816,446), PM-CID (11,339,676), SIMBAD (8,159,240) and VIAF (6,050,830).[6]

Actually, a major role of Wikidata as a hub for identification in the semantic web is recognised by the VIAF. In fact, VIAF uses Wikidata as an 'other data provider', i.e., a provider of data other than a National bibliographic agency. Among the Wikidata items with a VIAF identifier, most common identifiers registered in the personal items are, in decrescent order: ISNI (1,136,260; 18%); DBN (1,012,493; 16%); LC (983,206; 15%); NTA (480,580; 7%); SUDOC (431,919; 6,8%) and BNF (428,792; 6,8%) (Bargioni, Bianchini, and Pellizzari 2021, table 5). The relationship between Wikidata and VIAF is very strong. Wikidata uses property constraints to discover possible inconsistencies in statements both within Wikidata and in the external sources.[7] So, Wikidata users can check the issues and try to fix them, but any external service can take advantage of this characteristic too.

Identification in Wikidata is a process oriented to the quality of data. First, Wikidata explicitly requires – with the second notability criterion – that each item refers to "a clearly identifiable conceptual or material entity. The entity must be notable, in the sense that it can be described using serious and publicly available references".[8] For example, notability prevents Wikidata from accepting isolated clusters formed by VIAF based on a single contributor identifier. Second, clusters of identifiers in a Wikidata item are created by common users and not by automatically performed matches. Matches may be performed semi-automatically (by means of tools such as OpenRefine[9] or Mix'n'match[10]) but human control is always required. Moreover, as in authority work, references are mandatory for each triple and reference sources include encyclopedias, biographical dictionaries, scientific books and articles, in addition to VIAF and other national libraries authority data.

More and more Wikidata initiatives are oriented to improve the quality of authors' data. An example is offered by the bots: a bot in Wikidata "is a program that is allowed to upload large scale data and that is quality controlled by the community" (Siedlmayer 2020). For example, during SWIB 2020 OrcBot was presented: it is a tool created to take advantage of ORCID ids to improve the recording of the property that links the author items to their respective papers, based on ORCID Ids. The Enhancing author items process and issues of reconciliation between ORCID and Wikidata were the focus of the talk "Author items in Wikidata" at the WikiCite Virtual conference 2020 by Simon Cobb – wikimedian in residence at the National Library of Wales.[11]

3. Description in Wikidata

The "traditional" bibliographical description, marked by descriptive areas in ISBD, and fields and subfields of the MARC format, was firstly challenged by the birth of electronic catalogs, or rather

---

[6] https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all, visited 15 December 2020.

[7] Wikidata helps in identifying issues by two approaches: unique value violations and single value violations. A detailed description of both the approaches and their practical relevance as a quality control tool applied to VIAF is available in (Bargioni, Bianchini, and Pellizzari 2021).

[8] https://www.wikidata.org/wiki/Wikidata:Notability

[9] https://openrefine.org/.

[10] https://mix-n-match.toolforge.org/. See also (Agenjo-Bullón and Hernández-Carrascal 2020).

[11] https://upload.wikimedia.org/wikipedia/commons/7/79/Author_items_in_Wikidata.pdf

by the new, previously impossible, opportunity to search in any part of the description. Moreover, the electronic catalog challenges the need for a layout made up of descriptive areas, because the flag format of visualization, highlighting the metadata/data structure, overcome the value of the order of citation and the semantics of punctuation.

With the evolution of electronic catalogs and the gradual occurrence of major commercial players (at this stage) into the world of libraries and catalogs, we have therefore come to have tools that allow research and access to different databases, produced by different parties with different purposes. But in this situation the standardization of description is progressively fraying and being lost, in favor of a greater speed in the availability of information about resources, often produced directly by those who produce and make available the resource themselves, whether in analog or digital format.

However, this advantage is detrimental on the search side, as it increases the noise in catalogs and discovery tools, and an insufficiently skilled user may find it difficult to disentangle the results obtained. The impetuous technological evolution of the 21st century, together with a reflection on the functions and object of cataloguing that has led to radical changes in the way we approach resources, is beginning to show the consequences of all this in the cataloguing world. The semantic web and linked data are influencing the ways in which bibliographic data (in the broadest sense) are created, shared, and made available to potential users. In this phase, moreover, no-profit stakeholders outside the world of libraries are becoming increasingly present. An example is Wikidata, created and implemented by a community of volunteers with different training and mindset than those traditionally linked to the book professions, and by volunteers who deal with bibliographic data management.

On the side of libraries, traditional standards are losing their central role in bibliographic description. The static and linear MARC format, subject to many criticisms for many years, is giving way, with difficulty, to different models, such as BIBFRAME, which offers greater flexibility in line with the developments of the semantic web and linked data.

The ISBD format, still used as a standard for describing resources in some cataloging standards, is marking time for the moment. It remains the basis for both the practice and the teaching of cataloging in some settings and situations, but it cannot be denied that we are moving towards other newer and wider ways of describing, like RDA (Resource Description and Access).

RDA, in its first version still linked to AACR2, despite its innovativeness; but, in the new official version from December 2020 has radically changed the way to approach the description of resources. It uses IFLA LRM as a basis for its implementation, with some adaptations that the editors have considered essential for the "practical" needs of the library community, in line with what is expressed in the model, namely that implementations and changes are possible while respecting the basic structure.

On the one hand, RDA allows different levels of description with different types of data encoding (from the mere literal transcription to the use of IRI), on the other hand, it enables libraries using or wanting to use it to adopt different possibilities of use and implementation, with the only constraint to remain "faithful" to the framework and the general choices proposed by the standard and therefore to be interoperable with other realities that use it.

Anyway, an ethical problem must be highlighted: ISBD is a free descriptive standard, while RDA is not; if ISBD disappears, what will remain to those who cannot afford access to RDA?

The description in Wikidata, on the other hand, although conceptually very similar to the traditional one, presents differences and potentialities that a traditional description does not have and cannot have.

JLIS.it

As said, conceptually the basic elements of a description are those we are used to in the world of libraries, but we can immediately highlight some important innovations to increase the potential of the description. First, the full implementation of the modelling of the bibliographic universe of IFLA models, with the representation of Works, Expressions, Manifestations, and Items. Secondly, the possibility offered by Wikidata to qualify data. Finally, the possibility of integrating in the description identifiers of different types coming from different sources. While in traditional bibliographic description data qualification was impossible, in Wikidata this is not only feasible, but advisable. In this way you can achieve great advantages for all types of resources that you want to describe. Qualifying is not just specifying the data sources, but their chronological or geographical context; for example, the period of use of a form of a printer's name, a form of a place name, or the language used is a major advantage and a potential that has yet to be fully exploited.

Another aspect relevant to the concept of UBC is the possibility of going in depth in the description of resources. By this we mean that if the UBC very often stopped at the monographic level, in Wikidata instead it is possible to find descriptions of journal articles, or "sheets" of conference proceedings or miscellany. Indeed, perhaps because this lack is significant in traditional catalogues, these types of resources represent a very high percentage of the items in Wikidata.

Certainly, some aspects need to be improved, such as the correct attribution of properties to the right level and the creation of relationships, for example, between works, expressions, events, and items, but the potential is great and the foundations are sufficiently solid to be able to think of continuing the construction of a valid tool for a new vision of UBC, not tied to national conditioning or commercial logic.

The challenges that will have to be faced, at another level, will instead be those related to the use, and the visualization/reuse of these data, but this is not the place to delve into the matter.

Fig. 5. Example of a Wikidata description of a scholarly article. https://www.wikidata.org/wiki/Q58379188

JLIS.it

Fig. 6. Wikidata template with all the properties for the bibliographic description – sample; cfr. https://www.wikidata.org/wiki/Template:Bibliographic_properties

Fig. 7. Wikidata template with the Work item properties; https://www.wikidata.org/wiki/Wikidata:WikiProject_Books#Work_item_properties

JLIS.it

4. Wikidata as a bibliographic tool. Strategies, projects, and tools

4.1 Identification

To understand how the identification process can be improved in Wikidata, it is necessary to distinguish two possible approaches: by identifiers and by properties of the items.

The most important tool of quality control for proper identification of items are identifiers.[12] In fact, every property associated with an external identifier is provided with constraint rules – usually, an external identifier must be associated with only one item and one item must have only one identifier per type. These rules are extremely important for bibliographic control. In fact, they allow to identify possible errors within Wikidata (e.g., a duplicated item), but above all, they show possible mistakes also within the sources of the external identifiers linked to a Wikidata item (e.g., when a duplication of external identifiers occurs in a Wikidata item).

An example of how it works can be useful to understand how much Wikidata can help in the identification work for persons. A quick check of the identifiers associated with "Ferruccio Battolini" shows that three distinct VIAF IDs and two distinct ISNI IDs are associated with the same person (figure 8a).[13] This example is relatively simple, but things get more complicated with classical authors (e.g., poetess Saffo; figure 8b).[14]
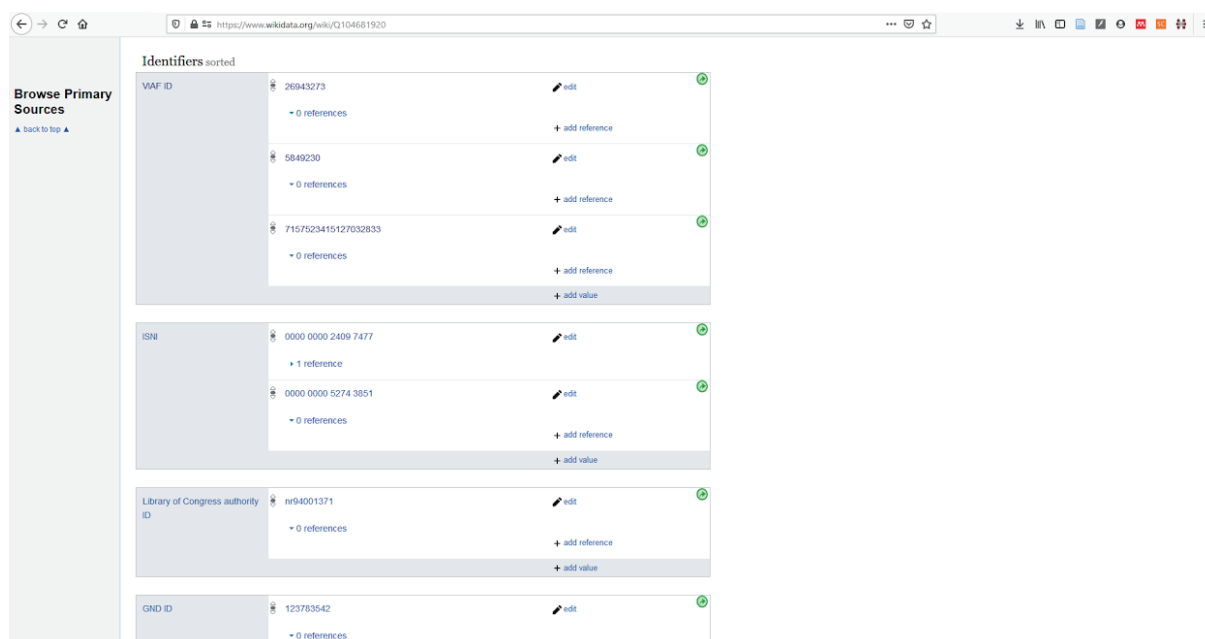

Fig. 8a. Duplicated VIAF and ISNI identifiers for Ferruccio Battolini

---

[12] See Wikidata Project: https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control.
[13] https://www.wikidata.org/wiki/Q104681920.
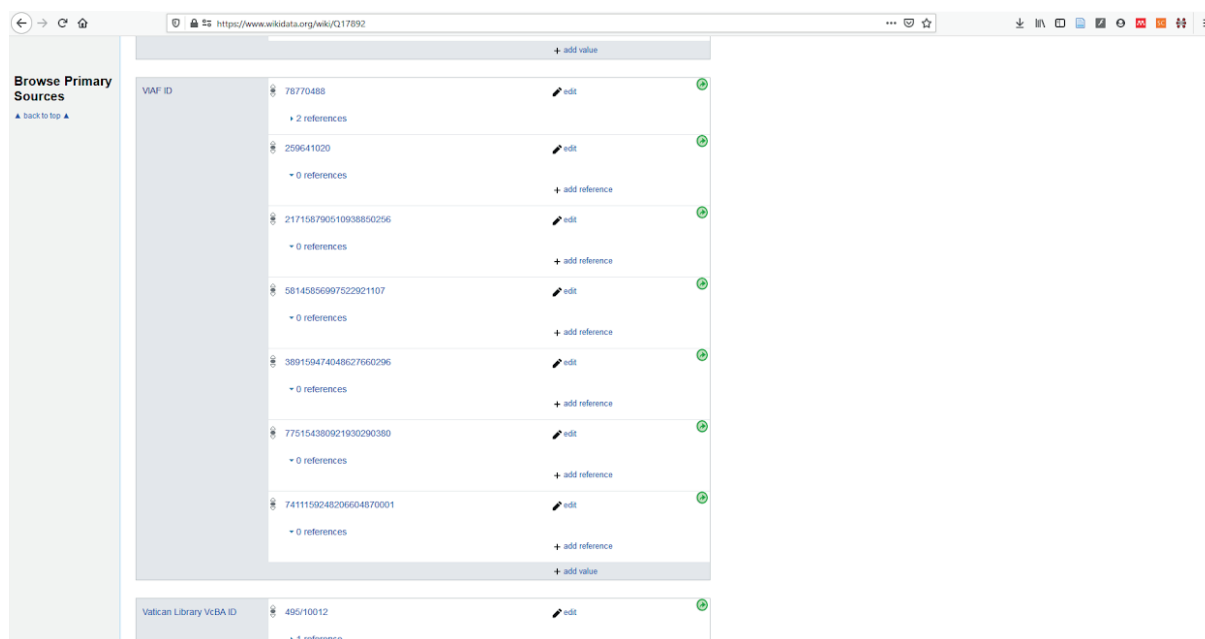[14] https://www.wikidata.org/wiki/Q17892.

JLIS.it

Fig. 8b. Duplicated VIAF identifiers for Sappho

As VIAF remains a major source for Wikidata, the community developed specific tools – named gadgets – to improve the reuse of its data in Wikidata items. Gadgets are enhancements of the edit interface for registered users and are very useful for data production.[15] For instance, the gadget *MoreIdentifiers* was created by Stefano Bargioni and Camillo Pellizzari to facilitate the creation of links between Wikidata items and VIAF entities and it enables users to add easily and quickly authority control IDs from VIAF with few edits checking the identifier and clicking on the button (figure 9).[16] Moreover, it enables to know whether an identifier is old or wrong (as it is presented strikethrough in red) and to create a report for any wrong identifier wrongly included in the VIAF cluster, if the case, by means of the thunder icon. A page of identifiers wrongly included in a VIAF cluster is maintained and constantly updated by Wikidata users; alas, it seems not so used by VIAF managers.[17]

---

[15] A list of gadgets is available at https://www.wikidata.org/wiki/Wikidata:VIAF/cluster#Gadgets.

[16] https://www.wikidata.org/wiki/User:Bargioni/moreIdentifiers.

[17] https://www.wikidata.org/wiki/Wikidata:VIAF/cluster/conflating_specific_entries.
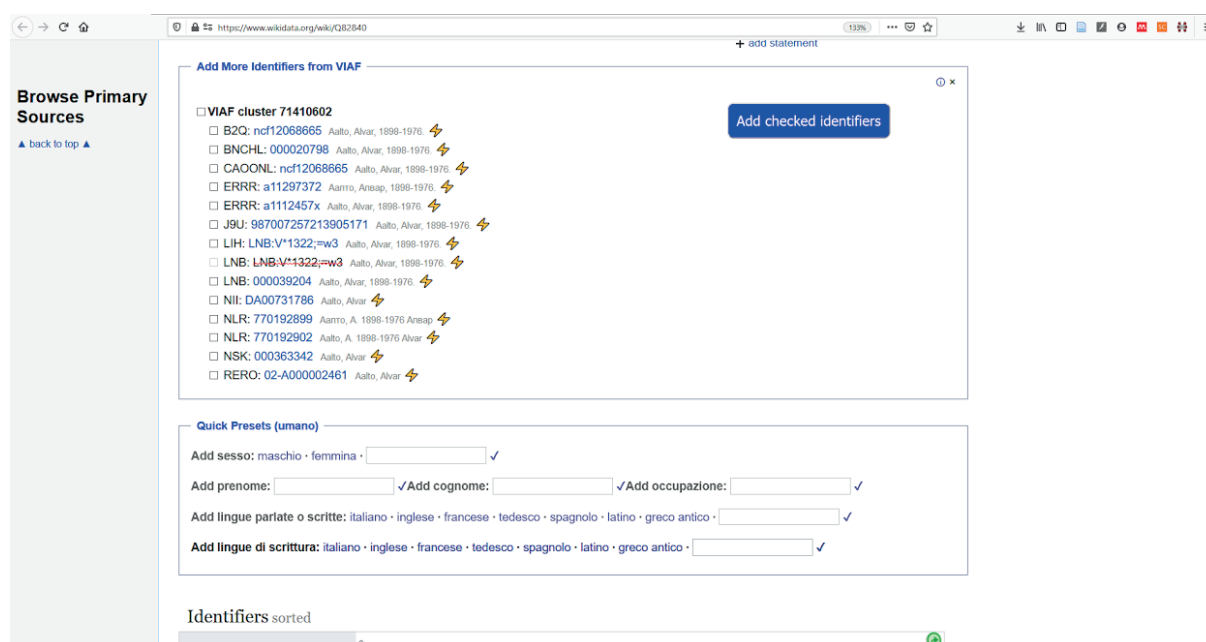
JLIS.it

Fig. 9. Box of the Wikidata gadget *MoreIdentifiers*.

MoreIdentifiers works for any kind of VIAF entity (such as geographic names or corporate names) but it is best useful for personal names.

Properties are key for the identification of items within Wikidata. In this case, identification is based on the matching of several properties. For example, human beings' identification is usually based on the matching of the label, the description, and the dates of birth and death.

In this approach, the more the available properties, the more the probabilities for identification. So, the number of properties of an item is a key issue, because a higher number of properties describing an entity assure a more probable identification – or disambiguation – of the two entities being compared.

For this reason, a dedicated gadget was developed by Wikidata community: *Recoin*, i.e., *Relative Completeness in Wikidata* (figure 10). Recoin is a "script that extends Wikidata entity pages with information about the *relative completeness* of the information" referring to the "extent of information found on an item in comparison with other similar items".[18] Recoin is a tool to help authors of Wikidata to know on which data attention must be focused on; moreover, it is also extremely useful for data consumers to be aware of the degree of information available about an item.

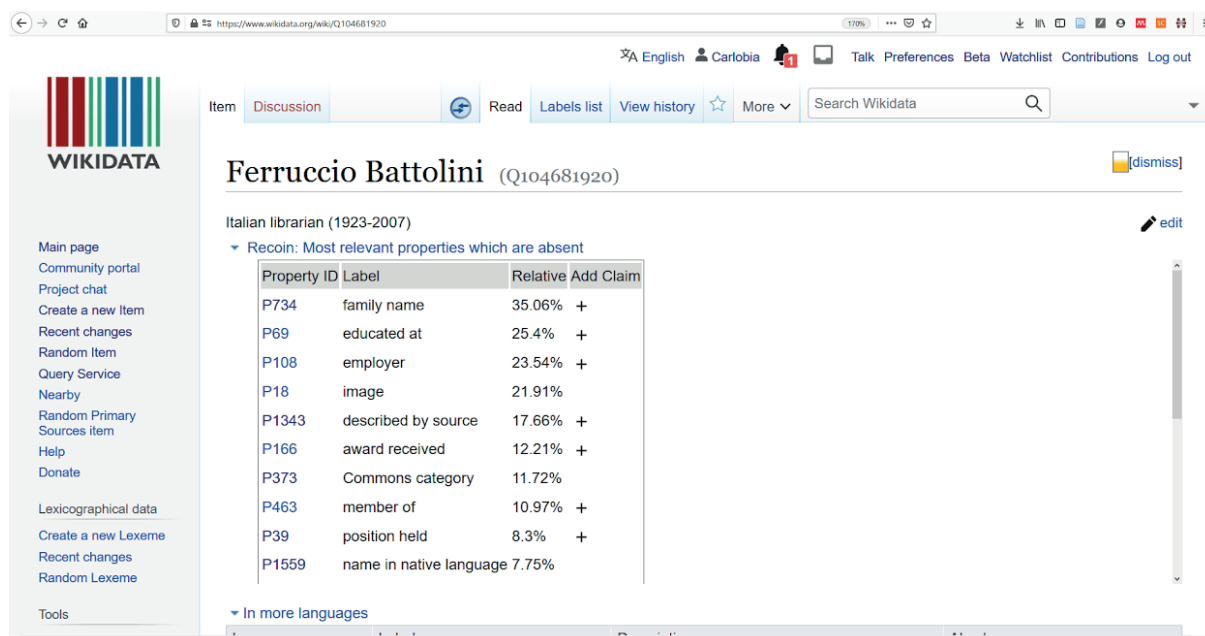18 https://www.wikidata.org/wiki/Wikidata:Recoin.

JLIS.it

Fig. 10. Example of missing properties highlighted by Recoin

As shown in figure 10, Recoin offers a status indicator icon, ranging from very detailed to very basic, to indicate the relative completeness of the description of an item on a 5-level scale, and a list of the most relevant properties which are not present in the item. Missing properties are detected by a comparison of the properties in that item and the properties most frequently occurring in that class. For example, the properties in an item representing a politician are compared to the most frequently occurring properties of the item belonging to the class 'politicians' (Balaraman, Razniewski, and Nutt 2018).

Nevertheless, identification within Wikidata is far from being perfect and can still be improved. Many new items are poorly described because of two main issues: many items are created by semi-automatic processes and, for this reason, data can be incorrect, generic (e.g., string versus author; cf. below),[19] or poor. In addition, at present Wikidata as a semantic web hub, is undoubtedly more oriented towards identifying than describing items.

When data derived from external sources are incorrect, their limit is inherited in the Wikidata item description (as seen above with VIAF identifiers). For example, a large part of the item creation work from sources like ORCID is made by bulk upload from bots; this means that in these cases "errors can persist for many months without being rectified and can be replicated in bulk editing of the description without detection" (Cobb 2020, 3).

External data can result in generic data too. For example, it is possible to import data from Zotero – a Reference Manager Software – to Wikidata, but the authors of the books or the articles are recorded as a *string of characters* (P2093), instead of a relationship between the item and the *author* (P50). And this happens with many other automatic tools, so that about

---

[19] https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all.

JLIS.it

135 million of authors are recorded as strings compared to just 20 million recorded as author relationships.

Poor external source data can produce a low number of average statements, and this means a poorer description and a more difficult process of identification of the items (mainly towards other external sources). For example, a study by Simon Cobb shows that items having an ORCID have a low number of statements; so that "the latest author items are sparse in comparison to older items, which have had longer to attract the curatorial efforts of community members" (Cobb 2020, 3).

Anyway, as author item relations allow for much richer analysis, to fix the issue, the special tool *Author disambiguator*[20] was created: it is a tool for editing the authors of works recorded in Wikidata and for assisting in converting "those strings into links to author items as efficiently and easily as possible".[21]

Another issue for authority control in Wikidata is that the data harvesting process is not structured; initiatives to upload data are very much, and sometimes based on semi-automatic tools, but there is not a clear overall design nor strategy, as typical of bottom-up approaches.

At the end of his analysis, Simon Cobb suggests a few steps to improve identification process for Wikidata items that can be applied to any kind of item and of external source; major suggestions are:

- Seek community consensus on minimum acceptable standard for author items created by bot imports.
- Define author data requirements for a variety of use cases.
- Review and validate data in existing author items.
- Organise an online workshop to facilitate discussion and collaboration between interested members of the Wikidata editor community and other stakeholders within and outside the Wikimedia Foundation projects.
- Establish a WikiProject special interest group (SIG) to focus on the improvement and maintenance of author items" (Cobb 2020, 10).

4.2 Description

The improvement of description in Wikidata can be approached from at least two points of view: the description of a remarkable variety of items in Wikidata, and the more traditional description of bibliographic resources.

In the first case (item description) the issues are certainly more intertwined with the problems of identification, because to have good descriptions, a correct identification of the entity is necessary and therefore the number and quality of properties necessary and sufficient for the description itself must be defined.

In the second case, instead, the improvement would require the growth of the available resources and their univocal identification when possible (by means of identifiers such as ISBN, ISSN, DOI). Problems arise for all the resources that do not have an identifier univocally assigned by an internationally recognized agency, but that have only identifiers assigned by the world of libraries (BID; etc.).

---

[20] https://author-disambiguator.toolforge.org/ .
[21] https://www.wikidata.org/wiki/Wikidata:Tools/Author_Disambiguator. See also (Smith 2020)

JLIS.it

A key issue is the quality of the source metadata, as digitized resources or "digital libraries" show when collecting the product of digitization from different sources (one example for all, Internet Archive). In such situations, the critical issues concern the description of the "less standardized" events and the need of a proper identification of the expressions and works on the one hand, and on the other, a trustful reconstruction of the physical presentation of the events, to facilitate more specialized or bibliographic research.

The description approach provided by RDA, i.e., based on identifiers and IRIs, is particularly effective in a context like Wikidata, and could lead to a significant growth of the number of described resources, and to an enhancement in the quality of the descriptions, as well as to the correct identification of the various entities.

Compared to this, the advantages related to the possibility of making a more granular and detailed bibliographic control than library catalogs are certainly notable (articles, miscellanea perusal, etc.); the possibility of inserting identifiers related to the catalogs of the major libraries or library systems worldwide also allows to satisfy the user function *to obtain*, which is often what most users want as a result of a search.

Finally, we remember that description and identification issues are inevitably intertwined.

5. Suggestions from Wikidata for the UBC

Wikidata is a Wikimedia tool to meet the needs of Wikimedia platforms, but it has relevant bibliographic features that can help to better understand the future of the Universal Bibliographic Control. In fact, Wikidata offers a completely new approach to data management that involves the way in which our community thinks and operates the Universal Bibliographic Control, both from a practical and theoretical perspective.

Wikidata is not designed as a bibliographic tool, and it is not oriented, nor limited, to bibliographic resources. For this reason, even if a data schema is available as Wikidata property page for works, editions, scientific articles, serials and so on, the quality and completeness of bibliographic data are usually high, but not certain. In fact, the number and the quality of the identifiers and properties recorded in Wikidata items are very varying, and the oldest items are usually more well-structured than the most recent ones; anyway, many gadget and tools (*MoreIdentifiers*, *Recoin*, *Author Disambiguator*, etc.) are available to improve them. Furthermore, while its bottom-up approach is a major asset in a global environment in which the role of great national bibliographic agencies is unable to fulfil the requirements of UBC, it is also a limit for the lack of a clear overall strategy of implementation of authority and bibliographic data.

Anyway, from a practical perspective, Wikidata is a clear example of the need for a new approach to identification and description, that are intertwined. First, a change in the workflow and in the mindset is required to the cataloguer, because a basic and even problematic identification must precede the description of the item.

Secondly, the relevance of globally preferred and variant access points is lessened; in fact, they remain relevant just in a local environment, and in a specific context defined by a particular set of rules. While labels and aliases pragmatically meet the requirements of making data accessible for any users' search, the identification function – a pillar of UBC – is assured by international

JLIS.it

identifiers, among which Wikidata ID is more and more significant. Moreover, the description in Wikidata – although conceptually very similar to the traditional one – presents differences and potentialities that a traditional description does not have and cannot have. First, the full implementation of the modelling of the bibliographic universe of IFLA models is available, with the representation of Works, Expressions, Manifestations, and Items. Third, the possibility of integrating in the description identifiers of different types coming from different sources, above all from the library field. Last, but not least, the possibility offered by Wikidata to qualify data. For instance, the chance to specify the period of use of a form of a printer's name, or of a place name, or of the used language is a major advantage and a potential that has yet to be fully exploited.

Another relevant point in Wikidata practical perspective is its major value as a *de facto* infrastructure to support the efficient exchange of bibliographic data among users, especially those who are not national bibliographic agencies. Wikidata is already a major hub of the semantic web also for bibliographic purposes. Moreover, Wikidata can record and disseminate bibliographic data of analytic descriptions, such as scholarly articles or chapters of books. Finally, Wikidata upgrades the concept of authority work, including reference both to the main international library catalogs and to local library catalogs and to a wider variety of reference sources (such as encyclopedias, dictionaries, and biographical repertories).

From a theoretical perspective, Wikidata offers a pragmatic way to think globally and act locally. For instance, it suggests looking at authority data as just a part of a wider perspective in which we produce and record data. For instance, it helps to recognize that an 'author' is just a person with a typed relationship toward a work, or a subject is any kind of entity with another typed relationship with a work. Authors and subjects, in a sense, do not exist in 'nature', but they become meaningful only in a bibliographic data perspective, and they must be expressed by a relation of authorship or aboutness between entities.

Furthermore, it shows that there is no need for one standardization of practices for establishing the headings and structure of authority records in one international form; instead, users' convenience can be achieved by a technological infrastructure capable to present to each user the information about an entity in its own language and script. Wikidata is the most evident example of the distributed and diffused approach of the semantic web to the issue of the universal identification of the entities. Moreover, thanks to a bottom-up and co-operative approach, Wikidata fulfils the requirements of International cataloguing Principles of common usage and convenience of the user by means of the users themselves.

There are other two relevant points about the contribution of Wikidata to the theoretical framework of the Universal Bibliographic Control. The first is that authority and bibliographic control must be tackled as just a part of the more general topic of the creation of a knowledge graph of all human knowledge by means of linked open data. In fact, Wikidata and the Semantic Web record data for any kind of item and not just for entities of bibliographic interest. In this new context, data for the achievement of the Universal Bibliographic Control and data, information, resources controlled by the Universal Bibliographic Control are perfectly integrated in one structure. The second is that this objective cannot be achieved only by contribution, cooperation, and networking of large National Agencies, as a larger number of stakeholders must be involved to achieve a UBC also including the full indexing of any kind of scientific communication.

JLIS.it

Bibliographic references

Agenjo-Bullón, Xavier, and Francisca Hernández-Carrascal. 2020. 'Wikipedia, Wikidata y Mix'n'match'. *Anuario ThinkEPI* 14. https://doi.org/10/ghbj6t.

Allison-Cassin, Stacy, and Dan Scott. 2018. 'Wikidata: A Platform for Your Library's Linked Open Data'. *Code4Lib Journal*, 4 May 2018. https://journal.code4lib.org/articles/13424.

Anderson, Dorothy. 1974. *Universal Bibliographic Control. A Long Term Policy - A Plan for Action.* Munchen: Verlag Dokumentation.

Association of Research Libraries. 2019. *ARL White Paper on Wikidata. Opportunities and Recommendations.*

Balaraman, Vevake, Simon Razniewski, and Werner Nutt. 2018. 'Recoin: Relative Completeness in Wikidata'. In *WWW '18 Companion: The 2018 WebConference Companion*, April 23–27, 2018, Lyon, France. New York, NY, USA: ACM. https://doi.org/10.1145/3184558.3191641.

Bargioni, Stefano, Carlo Bianchini, and Camillo Pellizzari. 2021. 'Beyond VIAF. Wikidata as a Complementary Tool for Authority Control in Libraries'. *Information Technology and Libraries* 40 (2). https://doi.org/10.6017/ital.v40i2.12959

Berners-Lee, Tim. 2006. 'Linked Data - Design Issues'. 27-7-2006. 2006. http://www.w3.org/DesignIssues/LinkedData.html.

Cobb, Simon. 2019. 'Connecting Persistent Identifiers in Wikidata'. In *Portland PID Workshop, 6th May 2019.* https://upload.wikimedia.org/wikipedia/commons/8/82/Connecting_persistent_identifiers_in_Wikidata.pdf.

———. 2020. 'Author items in Wikidata'. Presented at the WikiCiteVirtual Conference, October 26. https://upload.wikimedia.org/wikipedia/commons/c/cc/WikiCite_Virtual_Conference_2020_-_Author_items_in_Wikidata_-_Slides.pdf.

Dunsire, Gordon, and Mirna Willer. 2014. 'The Local in the Global: Universal Bibliographic Control from the Bottom Up'. In *IFLA WLIC 2014.* Lyon, France. http://library.ifla.org/817/.

Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Knudson Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, et al. 2020. 'Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage'. OCLC. 4 May 2020. https://doi.org/10.25333/faq3-ax08.

Gorman, Michael. 2014. 'The Origins and Making of the ISBD: A Personal History, 1966–1978'. *Cataloging & Classification Quarterly* 52 (8): 821–34. https://doi.org/10.1080/01639374.2014.929604.

Hernández-Cazorla, Iván, Manuel Ramírez-Sánchez, and Gregorio Rodríguez-Herrera. 2019. 'Wikidata, WikiCite y Scholia Como Herramientas Para Un Corpus de Datos Bibliográficos Enlazados. Curación y Estructuración de La Producción Científica de Los Investigadores Del IATEXT'. *PRISMA.COM* 40 (2019): 78–87.

IFLA Cataloguing Section and IFLA Meeting of Experts on an International Cataloguing Code. 2016. *Statement of International Cataloguing Principles (ICP).* Den Haag: IFLA.

JLIS.it

Illien, Gildas, and Françoise Bourdon. 2014. 'A la recherche du temps perdu, retour vers le futur: CBU 2.0'. In *IFLA WLIC 2014*. Lyon, France. http://library.ifla.org/956/.

Lemus-Rojas, Mairelys, and Jere D. Odell. 2018. 'Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project Using Wikidata and Scholia'. *Journal of Librarianship and Scholarly Communication* 6. https://doi.org/10.7710/2162-3309.2272.

Linked Data for Production. 2020. 'Wikidata as a hub for identifiers'. Google Docs. 11 June 2020. https://docs.google.com/presentation/d/1jWz3_nCf5rdd-7ejETGlfv99UV2PnD1v/edit?usp=embed_facebook.

Mietchen, Daniel, and Lane Rasberry. 2020. 'Presenting Scholia. A Scholarly Profiling Tool'. Presented at the LD4 Wikidata Affinity Group, August 11. https://docs.google.com/presentation/d/1jJbYSnYSDh36-LxzSpedFyWUzusZAjuBbP-y46ji-0w/edit#slide=id.g35f391192_00.

Nguyen, Ba Xuan, Jesse David Dinneen, and Markus Luczak-Roesch. 2020. 'A Novel Method for Resolving and Completing Authors' Country Affiliation Data in Bibliographic Records'. *Journal of Data and Information Science* 5 (3): 97–115. https://doi.org/10/ghsnkn.

Nielsen, Finn Årup, Daniel Mietchen, and Egon Willighagen. 2017. 'Scholia, Scientometrics and Wikidata'. In *The Semantic Web: ESWC 2017 Satellite Events*, edited by Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and Olaf Hartig, 10577:237–59. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-70407-4_36.

Seidlmayer, Eva, Jakob Voß, Tetyana Melnychuk, Lukas Galke, Klaus Tochtermann, Carsten Schultz, and Konrad Forstner. 2020. 'ORCID for Wikidata – Data Enrichment for Scientometric Applications'. In *Proceedings of The 1st Wikidata Workshop*. https://wikidataworkshop.github.io/papers/Wikidata_Workshop_2020_paper_9.pdf.

Smith, Arthur P. 2020. 'Author Disambiguation'. In *WikiCite 2020 Virtual conference*. https://upload.wikimedia.org/wikipedia/commons/3/38/WikiCite_2020_Author_items.webm.

Unesco/LC Bibliographical Survey. 1950. *Bibliographical Services: Their Present State and Possibilities of Improvement*. Washington: Library of Congress.

Veen, Theo van. 2019. 'Wikidata: From "an" Identifier to "the" Identifier'. *Information Technology and Libraries (Online)* 38 (2): 72–81. https://doi.org/10/ghbj62.

Vrandečić, Denny, and Markus Krötzsch. 2014. 'Wikidata: A Free Collaborative Knowledgebase'. *Communications of the ACM* 57 (10): 78–85. https://doi.org/10/gftnsk.

W3C Incubator Group. 2011. 'Library Linked Data Incubator Group Final Report'. http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.

JLIS.it

"Discoverability" in the IIIF digital ecosystem

Paola Manoni(a)

a) Biblioteca Apostolica Vaticana, https://orcid.org/0000-0001-7802-2718

ABSTRACT

The IIIF APIs have been used since 2012 by a community of research, national and state libraries, museums, companies and image repositories committed to providing access to image resources. The IIIF technical groups have developed compelling tools for the display of more than a billion IIIF-compatible images.

We can figure out that with hundreds of institutions participating worldwide, the possibilities, for instance, for IIIF-based scholarship are growing so one question could be about the discovery of those images relevant to one's research interests in order to discover them for their consultation or, even more, for their reuse.

While IIIF specifications discussion has focused on the machine-to-machine mechanisms of making IIIF resources harvestable, we have yet to implement an end-to-end solution that demonstrates how discovery might be accomplished at scale and across a range of differing standards for metadata arising from libraries, archives, and museums.

KEYWORDS

IIIF; LAM; Discovery; Digital ecosystem.

JLIS.it

1. Discoverability

The International Image Interoperabilty Framework[1] as an interoperability protocol for image resources held in libraries, archives, museums, has produced over a billion IIIF-compliant images. This paper will focus on how this vast production is actually changing not only the use of digital objects online in the context of tools at the convenience of digital humanities, for instance with the well known abilities of IIIF viewers, such as Mirador[2], but also the concept of discoverability of the knowledge objects now available via IIIF.

Discoverability is the quality of being able to be discovered or found and in relation to online content, it is the quality of being easy to find via a search engine, within an application, or on a website.

If we focus on the discoverability in the IIIF context we can refer to two main aspects:
- Which are the requirements that make a web platform a discoverable digital library service in the light of IIIF;
- How is it possible to discover IIIF-compliant content through current web platforms.

A first look about the context of the two issues, is concerning the non-trivial meaning of LAM data in the universe of a single domain, for an evaluation of their impact on the discoverability of IIIF objects. We thus consider the abstraction of LAM data produced within the digital ecosystem starting from the traditional statements related to the classes of:
- *"Structured data* – In the LAM domain they include bibliographies, catalogs, indexing and abstracting databases, authority files. Structured data is generally stored in databases where all key / value pairs have clear identifiers and relationships and follow an explicit data model
- *Semi-structured data* - they are the unstructured sections within metadata descriptions as well as any unstructured portions of structured datasets.
- *Unstructured data* – they are the typical "everything else" pertaining to documents and other information-bearing objects in all kinds of formats." (Zeng 2019).

We consider the typical elements of the IIIF Presentation API, keeping in mind that IIIF Presentation API provides:
- A model for describing digital representations of objects: just the metadata chosen in a completely arbitrary way in order to offer a remote viewing experience.
- A format for software - viewing tools, annotation clients, web sites - to consume and render the objects and any other associated content in the form of annotations.

This does not mean that descriptive metadata has no place in a digital object provided by the Presentation API. In fact, it is important that the object is linked to its description and to all the information relating to it. The presentation API provides this human readable information, so that viewers can interpret the important contextual information to end-users.

---

[1] Cfr. *International Image Interoperability Protocol* < https://iiif.io/>. Accessed April 15, 2021.

[2] Cfr. https://projectmirador.org/. Accessed April 15, 2021. Mirador is a fully IIIF-compatible tool capable of interpreting IIIF APIs. Mirador is an open source image, Javascript and HTML5 viewer that delivers high resolution images in a workspace that enables image annotation and comparison of images from repositories dispersed around the world, starting from compatibility with Image API that specifies a web service returning an image in response to a standard HTTP or HTTPS request.

JLIS.it

The information pertaining to Presentation API is the IIIF manifest of the digital object represented as a "thing" and enriched with the complex knowledge data related to it.

A so-called IIIF manifest contains a descriptive section of the digital object but the specifications do not define any rules relating to metadata. In other words, we can say that IIIF requirements are completely agnostic as to which descriptive metadata to apply as well as to which image formats.

The galaxy of data pertaining to the "thing" represented in the manifest is completely scalable and referable to the different meanings of LAM data (structured, semi-structured, and unstructured) we mentioned.

Moreover, as for the Presentation API, the meaning of any accompanying descriptive metadata for display in a viewer is not taken into any account. The purpose of this API is the representation of the content of the work – for example the pages of the book, the painting – or the link where users can get information about the meaning of the content of the work.

The objective of the IIIF Presentation API is to provide the information necessary to allow a rich, online viewing environment for primarily image-based objects to be presented to a human user, likely in conjunction with the IIIF Image API. In other words, the IIIF Presentation API gives us a specification for "presenting" a digital object and the data describing it in order to view, annotate it, or compare it with other objects. A IIIF client can also display any accompanying metadata included as pairs of labels and values within the manifest. But it needs no definition or scheme for what that metadata means. It is *outside of the scope* of the Presentation API (Crane 2017).

The user can view important semantic metadata, but the scope of the Presentation API is just to leverage that text. In the Presentation API, the semantic meaning is *elsewhere* because it is not belonging to its the specifications. An API client should simply render them.

In a nutshell, a manifest is what a IIIF viewer loads to display the object. A manifest could be used to represent the object within a web service as well as it could be used to add annotations to the represented object or even to be aggregated within a new manifest thus realizing its reuse.

The structure of the manifest also includes the concepts of sequence, which is of fundamental importance for aggregated resources (e.g. books, manuscripts and archive materials composed of page, leaf, folio or sheet) and canvas.

Each view of the object, for example each page is represented by a canvas. A Manifest contains one or more **Sequences** of **Canvases**. But a canvas is not the same as an image. "The canvas is an abstraction, a virtual container for content" (Crane 2017).

> A Canvas is the digital surrogate for a physical page which should be rendered to the user. Each Canvas has a rectangular aspect ratio, and is positioned such that the top left hand corner of the Canvas corresponds to the top left hand corner of a rectangular bounding box around the page, and similarly for the bottom right hand corners. The identifier for the Canvas is not an identifier for the physical page, it identifies the digital representation of it.[3]

The canvas is a kind of conceptual extra layer in which an object is included.
The canvas keeps the content separate from the conceptual model of the page of the book, paint-

---

[3] Cfr. *Shared Canvas Data Model* <http://iiif-io.us-east-1.elasticbeanstalk.com/model/shared-canvas/>. Accessed April 15, 2021.

JLIS.it

ing or archival unit. The content, we are referring to, could be blocks of text, videos, links to other resources, and it is exactly mapped on the canvas. By including a canvas in a manifest, you provide a space on which users and scholars can annotate the content.

All association of content with a canvas is done by **annotation**. The IIIF Presentation API is built on the W3C Web Annotation Data Model[4].

> Annotations associate content resources with Canvases. The same mechanism is used for the visible and/or audible resources as is used for transcriptions, commentary, tags and other content. This provides a single, unified method for aligning information, and provides a standards-based framework for distinguishing parts of resources and parts of Canvases.[5]

The canvas establishes a stage in which the simplest case – one image per canvas – is straightforward, but more complex cases, more complex and interesting associations of content, can be managed.

The latest specification of the IIIF, still in beta version, is the IIIF Content State API[6] which demonstrates another purpose for a representation by sharing a content to be represented on a canvas.

In its scope there are two examples:

- A user follows a link from a search result, which opens a IIIF viewer. The viewer focuses on the relevant part of the object, such as a particular line of text that contains the searched-for term.
- A user opens several IIIF Manifests to compare paintings, then wishes to share this set of views with a colleague.

These are examples of sharing a resource, or better, a *particular view* of a resource. Other examples include bookmarks, citations, playlists and deep linking into digital objects.

The objective of the IIIF Content State API is to provide a standardized format for sharing of a particular view of one or more IIIF Presentation API resources, such as a Collection, a Manifest, or a particular part of a Manifest.

Content State API **is how we can point at things in IIIF** and this demonstrates how the concept of digital resource and its reuse expand to include new ways of knowing the resources and new ways of citing them. In fact, it basically means dereferencing URIs of annotations whose motivation such as *content state* will be included in a manifest.

> Content State is a way for humans to share bookmarks, and it's also a way for search results to point at the exact part of a digital object that they match (Crane 2021).

We can argue at this regard that the semantic enrichment process pertaining to the IIIF's vision of LAM objects and data reflects the broader general transformation from document-centric to entity-centric knowledge modeling due to the many relations for each canvas.

---

[4] Cfr. *Web Annotation Data Model*. Accessed April 15, 2021. https://www.w3.org/TR/annotation-model/. The Model does not prescribe a transport protocol for creating, managing and retrieving annotations. Instead, it describes a resource oriented structure and serialization of that structure that could be carried over many different protocols.

[5] Cfr. *IIIF Presentation API 3.0* Accessed April15, 2021. https://iiif.io/api/presentation/3.0/.

[6] Cfr. *IIIF Content State API 0.3* Accessed April 15, 2021. https://iiif.io/api/content-state/0.3/.

JLIS.it

Let us now go back to consider the discoverability of the IIIF in the light of Presentation API and the first question

- Which are the requirements that make a web platform a discoverable digital library service in the light of the IIIF;

We may focus on this by considering the use case of the Vatican Library as an example.

2. The use case of the Vatican Library

DVL (the DigiVatlLib, <https://digi.vatlib.it>) is a digital library service. It provides free access to the Vatican Library's digitized collections: manuscripts, incunabula, archival materials and inventories as well as graphic materials, coins and medals, printed materials. It is fully based on the International Image Interoperability Framework technology, making digital materials easily accessible and usable.

- The viewer is able to zoom, browse and 'turn pages' of JPEG2000 images as well as allow scholars to compare digital objects from different IIIF repositories of other digital libraries.
- Descriptions and bibliographic references from the online catalogues are indexed and linked to digital materials.
- Each object is equipped with URIs for the discovery of IIIF manifests.
- The guided navigation ('faceted search') leverages metadata elements for narrowing or refining queries.

The Library has promoted a *new* perspective to the study of manuscripts by means of web communication and IIIF.

To meet this challenge the Library has implemented a project to enrich the digital delivery of these materials by annotating some exemplary manuscripts with scholarly analysis.

The use case of annotations in IIIF was a three-year Mellon-funded project, held between 2016 and 2019, in conjunction with Stanford University Libraries, which produced over 26,000 annotations for a selection of manuscripts chosen in the context of thematic pathways. In this platform (available at: <https://spotlight.vatlib.it>) the content of all the annotations is indexed along with the metadata, thus constituting a semantically enriched system that allows scholars to query an integrated search of all the available contents of a resource.

The project aimed to demonstrate, among the advantages of the IIIF for manuscripts, how the annotation level is a fundamental innovation for the study of contents: transcriptions, comments, comparative analysis of texts and images.

Thanks to the funds received, the Library has implemented a workflow using Mirador with scholarly analysis in order to tell scholarly narratives.

The Vatican Library has intended to engage the visitors to its website on the possibilities for using annotated manuscripts in IIIF, according to specific themes, by providing tools for discovering and comparing digital materials.

The deep analysis of contents of manuscripts entails the understanding of the "pre-print" world in which the manuscript is born. This implies a knowledge pertaining to the history of the man-

JLIS.it

uscript, its origin, provenance as well as other circumstances of the production of a manuscript; identifications of dates, scribes, artists; discussions about the intellectual content and descriptive discussion on paleographic matters.

In its essential lines, a thematic pathway is composed by three different kinds of information:

- A general description (introduction, historical information, etc) of the chosen theme, it represents the "Story";
- Descriptive and structural metadata and a curatorial narratives for each manuscript;
- Annotations, comments, in-depth analysis about detailed parts of a manuscript (e.g. texts, comments, illuminations, etc.) and transcriptions of units of information.

The four thematic pathways

1. The first one is about *Courses in Paleography (Greek and Latin, from antiquity to the Renaissance)*

   The rich collection of manuscripts preserved in the Library makes it possible to follow the evolution of the Greek and Latin scripts all the way from antiquity to the Renaissance.

   The availability of on-line images of manuscripts, together with the possibilities offered by the IIIF APIs, allows a complete transformation of teaching practice in this field.

   For each of the sections (Greek and Latin) of this thematic path, a set of complete digitized manuscripts, chosen to illustrate the phases in the development of the script from the fourth to the sixteenth century, is provided. From each manuscript, chosen pages with a paleographical and codicological description and a diplomatic transcription is also made available.

2. The second one is about *The evolution and transmission of texts of specific works: Latin Classics*

   The Vatican Library owns one of the most important collections of manuscripts with texts by Classical Latin authors, many of them richly illustrated.

   The aim of this pathways is to describe 81 manuscripts directly from the original codices: metadata and annotations pertaining to the study of texts and illuminations have been provided. The work throws light not only on the illustrations of the texts but especially on the relationship between text, illuminations, comment and the gloss.

   The importance of this project lies in the remarkable variety of typologies of the Classical world.

3. The third one is about *Vatican Palimpsests: Digital Recovery of Erased Identities*

   The Vatican Library has identified more than 380 manuscripts in its own collections, which include palimpsests, erased and then recycled parchment folios. This pathway intends to present this rich and scarcely explored material to the public by making an in-depth archaeological research on the palimpsests of twenty-four select manuscripts and recover their lost identities with the help of IIIF technology.

   Making accessible hardly legible images to the public is a challenging task because the

JLIS.it

actual method of publication has been designed to typical objects. By the pathway, digital reconstruction makes four palimpsests accessible both by their upper and lower scripts, a condition which the actual conservation of these manuscripts and the normal method of publication do not allow.

Erased texts are often very old and significant witnesses of a lost past but they are difficult to access for the naked eye. They need an expert interpreter and highly special photographic and post-processing technologies and especially the flexibility of presentation offered by IIIF APls which can turn erased texts more accessible online than in their physical existence.

4. The last one is about The humanist prince's library: Federico da Montefeltro and his manuscripts

The library of Federico da Montefeltro, Duke of Urbino (since 1474), is known as a typical humanist collection.

The collection was outstanding not only for its substance (the amount of volumes as well as the quality, in relation with other libraries of that age), but for the value of each manuscript partly acquired from antique market, many commissioned by Federico and realized by refined copyists and greater artists of that time. The manuscripts were produced in two main locations: Florence and Urbino.

In the first years, Federico preferred to buy or order manuscripts in Florence (both in writing and in illumination), later he preferred Ferrara or Padoan artists and scribes active in Urbino.

This pathway points out the characteristics of the two schools, very different in style, and the most important artists (half of the chosen manuscripts is representative of the Florentine school while the other half of the Ferrara and Padoan schools).

3. IIIF Discovery for Humans Community Group

IIIF enables the creation of rich digital collections that bring together content distributed among cultural heritage institutions. With image viewers, one is able to analyze works held in physically different locations side-by-side or overlaid within a web browser. However, in order to take advantage of the research tools afforded by IIIF, a user must be able to find IIIF resources.

Interoperable objects are of no use if one cannot find them, particularly if relevant objects reside in servers in many different institutions. Discovery in this case means human searching, browsing and finding of IIIF resources across institutions. To be successful, IIIF discovery must be user-focused and meet defined users' concrete needs.

To meet these needs, the IIIF Discovery for Humans Community Group was recently organized. This group aims to go beyond specification work to promote implementations that enroll experts in research, content, user experience, metadata, and various technologies. In order to advance discovery in the LAM space, this group will foster user-focused approaches enabling the targeted discovery, spanning institutional and domain silos, of IIIF resources.

These aims are different from and complementary to the approach of the IIIF Discovery Tech-

JLIS.it

nical Specification Group, which is chiefly concerned about providing the technical means for locating and finding updates about IIIF resources, as a prerequisite for harvesting and indexing metadata for searching within and across these institutional collections.

If we focus again on the two questions arisen in this paper about:

- Which are the requirements that make a web platform a discoverable digital library service in the light of the IIIF;
- How is it possible to discover IIIF-compliant content through current web platforms.

We may say that both are of fundamental interest to this group and they are closely related to the purposes of the initiatives conducted by the Group, aimed at:

- Gather problem statements and use cases to understand needs for user-focused discovery of IIIF resources
- Develop specifications for metadata attributes and crosswalks to enable discovery of LAM IIIF content across institutions and domains
- Create and maintain a list of metadata profiles in use by IIIF-supporting institutions to promote consistency in semantic description and its consumption
- Frame small-scale experiments that work towards live discovery implementations
- Provide a venue for demonstrations of applicable discovery applications and technologies
- Maintain a registry of existing discovery efforts
- Build on and amplify the ongoing work of the Discovery Technical Specifications Group and the IIIF Technical Community
- Communicate and disseminate the work of the group to the larger IIIF community, as well as allied professional communities.

One of the recent activities of this group was to **collect a list of discovery features** in order to:

- Provide examples of features as implemented
- Extract a comprehensive list of discovery features
- Develop a feature typology
- Identify IIIF-specific discovery affordances
- Build a feature checklist for self-evaluation
- Inform development of other discovery platforms
- Showcase exemplary feature implementations

First of all this task has provided a collated list of discovery features "in the wild," to better understand the current landscape of IIIF resource discovery. It was useful as a basis for compiling a wide-ranging list of possible discovery features. This was further condensed and organized to derive broader categories for these features, and to identify which features were specifically tied to IIIF rather than associated discovery more generally. From this analysis we were able to get a sense of which features are broadly implemented and which are rarer. Based on this work we developed a feature checklist that may be used to identify which core discovery features are and are not available on a given site.

Metric for discovery features was the first important milestone planning the group's commitments as an important first step to face the second question: How is it possible to discover IIIF-compliant content through current web platforms, the thread underlying this paper.

JLIS.it

References

Crane, Tom. 2017. *An Introduction to IIIF.* Digirati. Accessed April, 15 2021. http://resources.digirati.com/iiif/an-introduction-to-iiif/.

Crane, Tom. 2021. *What is IIIF Content State?* Accessed April 15 2021. https://tom-crane.medium.com/what-is-iiif-content-state-dd15a543939f.

Manoni, Paola. 2020. "L'adozione del IIIF nell'ecosistema digitale della Biblioteca Apostolica Vaticana." *DigItalia* 2: 96-105.

Manoni, Paola - Ponzi, Eva. 2020. "Thematic Pathways on the Web: IIIF Annotations of Manuscripts from the Vatican Collections: il "Progetto Mellon" della Biblioteca Vaticana". *Rivista di Storia della Miniatura* 24: 211-216.

Salarelli, Alberto. 2017. "International Image Interoperability Framework (IIIF): a panoramic view". *JLIS* 8, no. 1. Accessed April,15 2021. doi: 10.4403/jlis.it-12090.

Zeng, Marcia Lei. 2019. "Semantic enrichment for enhancing LAM data and supporting digital humanities: Review article" *El profesional de la información* 28, no 1.

JLIS.it

Bibliographic Control of Research Datasets: reflections from the EUI Library*

Thomas Bourke[(a)]

a) European University Institute

ABSTRACT

The exponential growth in the generation and use of research data has important consequences for scientific culture and library mandates. This paper explores how the bibliographic control function in one academic library has been expanded to embrace research data in the social sciences and humanities. Library bibliographic control (BC) of research datasets has emerged at the same time as library research data management (RDM). These two functions are driven by digital change; the rise of the open science and open data movements; library management of institutional repositories; and the increasing recognition that data sharing serves the advancement of science, the economy and society. Both the research data management function and the bibliographic control function can be enhanced by librarians' awareness of scholarly projects throughout the research data lifecycle (input, elaboration and output) – and not only when research datasets are submitted for reposit. These library roles require knowledge of data sources and provenance; research project context; database copyright; data protection; data documentation and the FAIR Guiding Principles, to make data findable, accessible, interoperable and reusable. This case study suggests that by creating synergies between the research data management function (during research projects) and the formal bibliographic control function (at the end of research projects) – librarians can make an enhanced contribution to good scientific practice and responsible research.

KEYWORDS

Research data | Datasets | Research data management | Bibliographic control.

JLIS.it

1. Introduction

Research libraries have a long history of collecting, managing and providing access to data resources – in particular, statistical data series in support of the social sciences. While there are important differences between disciplines and sub-disciplines, most research libraries had a limited role in the management of their institutions' research data *outputs* until the 21st Century. Today, the collation, bibliographic control, preservation and dissemination of research data are important library functions, due to increasing awareness of datasets as 'first-class' outputs of research.

This case study treats the management and bibliographic control of research dataset outputs in the social sciences and humanities at the Library of the European University Institute (EUI).

While research data management (RDM) is primarily carried out by scholars during research projects, librarians have steadily increased their collaboration and training to fill the skills and capabilities' gap in this area. RDM is undertaken throughout the research data lifecycle, embracing the control of data inputs, the elaboration of data, the protection of data and the creation of research data outputs and documentation. The main reasons for librarians' involvement in research data management include: the exponential growth in the availability and production of digital data; the rise of the open science and open data movements; the establishment of research repositories – frequently managed by libraries; and the increasing recognition that the sharing of research data serves the advancement of science. All of the above are in contexts which require the transfer of knowledge and expertise of library staff – through consultation with, and training of, researchers. While librarians' research data management takes place *during* research projects; bibliographic control normally takes place *after* (or towards the end of) research projects.

Both research data management (Section 3 below) and the bibliographic control of datasets (Section 4 below) require librarians to strengthen their liaison with researchers in order to enhance familiarity with research project design, data generation and use, and research data outputs. Research data management also requires librarians to support the generation of data management plans (DMPs) which are increasingly required by science funders.[1] Support for data management planning raises librarians' awareness of the nature and scope of research projects before final data outputs are presented for reposit. While campus libraries have important roles regarding research data management and bibliographic control, it is acknowledged that – in some institutions – there are lead roles for data centres, ICT services and/or research administration offices.

2. Data, digital change and scientific culture

The generation, collection and use of vast quantities of data – and the retro-digitisation of non-digital collections and content – places research libraries at the vanguard of recent transformation.[2] In

---

[1] The European context is described by Filip Kruse and Jesper Boserup Thestrup (Kruse and Thestrup 2018).
[2] The evolution of library research data roles is analysed by Robin Rice and John Southall (Rice and Southall 2016); by Lynda Kellam and Kristi Thompson, et al. (Kellam and Thompson 2016); and by Rossana Morriello (Morriello 2020).

JLIS.it

addition to the volume of data, the tools for the elaboration of data have become more sophisticated. In the social sciences and the humanities, these developments have had an impact on scientific culture – facilitating more empirical and applied research; experimental research; evidence-based policy research, and data-driven methodologies.

The definition of 'data' varies across academic disciplines and sub-disciplines and the scope of the term itself has been debated for several decades.[3] Data types in the social sciences and humanities include: numerical data, minable text, survey data, experimental data, interview transcripts, archival material, field notes, images, and audio and video recordings. The long history of library expertise in the management of multi-media collections constitutes a solid basis for library curation of research data outputs.[4]

Although definitions vary by discipline, it is useful for librarians to distinguish between collected or acquired 'databases' (eg. databases of monographic, journal or statistical content) and individual 'datasets' (eg. data outputs from research projects hosted at their institutions). This paper does not treat the traditional library database acquisition and management function (which includes the acquisition, classification, cataloguing and access control of subscription resources; eg. financial market data).[5] The data treated in this case study are *outputs* generated by university scholars, which are managed, bibliographically controlled, curated, classified and reposited by library staff for the purpose of preservation and – where possible – sharing with other researchers.

The open data movement – an extension of the open access movement – refers to a growing trend whereby government agencies, international organisations and researchers share data outputs, documentation, codebooks and software via the internet. Here it is necessary to distinguish between 'public data' and 'research data.' Most governments and international organisations provide some level of access to 'open public data.' In the research community 'open research data' refers to outputs from scholarly research projects which are openly available, usually via institutional repositories.

3. Research data management: library roles

The impact of technological change on scientific culture has necessitated the expansion of library data support roles. The traditional function of acquiring access to subscription databases has been joined by two newer library roles: (i) library support for data-intensive research and research data management *during the research data lifecycle* and (ii) the bibliographic control, collation, reposit and preservation of research data outputs *at the end of the research project*. Research data management during research projects is carried out by both scholars and librarians, complementing their

---

[3] For a theoretical treatment, see the entry "Data" in the Encyclopaedia of Knowledge Organization: https://www.isko.org/cyclo/data (International Society for Knowledge Organization, n.d.).

[4] See Joudrey 2015, chap. 5; and Pradhan 2018.

[5] At the EUI, these resources are presented in the Library Data Portal: https://www.eui.eu/Research/Library/Research-Guides/Economics/Statistics/DataPortal and classified at: https://www.eui.eu/Documents/Research/Library/Research-Guides/Economics/Statistics/MacroMicroLocations.xls (Accessed 7 April 2021).

JLIS.it

respective expertise.[6] Bibliographic control, reposit and preservation of research data outputs are carried out by librarians. This is especially true when it comes to datasets in disciplines where the culture of managing (and sharing) research data is not yet fully developed, or where established subject-oriented data repositories (e.g. GenBank, HEPData) do not exist.
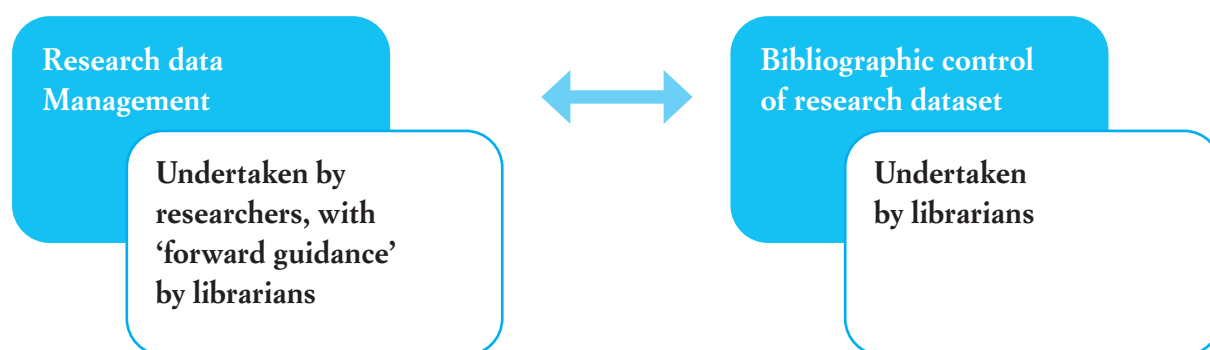

Fig. 1. RDM and BC roles: scholars and librarians

An important component of research data management is the generation of data management plans (DMPs).[7] A request from a principal investigator for DMP support is often the first point of contact between a librarian and a new research project. Data management plans provide information on how data is generated and/or sourced; how data is organised, used and elaborated; how data – and data subjects – are protected; how data and tools are described and documented; how data is stored and secured during the research project; how data authorship and credit are assigned; how data will be preserved and whether research data outputs can be shared.[8] The involvement of librarians in data management planning constitutes a solid foundation for the eventual bibliographic control of dataset outputs.

Contemporary research data management is underpinned by the FAIR Guiding Principles, to make data *findable, accessible, interoperable* and *reusable* – frequently used by librarians to promote awareness of good research data management practices.[9] Both the research data management function and the bibliographic control function help advance the FAIR Guiding Principles. Library research data management 'forward guidance' is provided via individual user support, library-web documentation and group training. Librarians provide advice that data outputs must be carefully structured, because the 'objects' (outputs) for reposit will be datasets (not unstructured data observations) and that researchers should carefully consider the design of datasets early in their research projects. Dataset structure varies by discipline and sub-discipline, medium, types

---

[6] The Consortium of European Social Science Data Archives (CESSDA) maintains an annually-updated Data Management Expert Guide. https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide (Consortium of European Social Science Data Archives, n.d.).

[7] Online template tools such as DMPonline https://dmponline.dcc.ac.uk/, maintained by the UK Digital Curation Centre (Digital Curation Centre, n.d.), and Argos https://argos.openaire.eu/splash/, maintained by OpenAIRE (OpenAIRE, n.d.), can be used to generate structured data management plans.

[8] https://www.eui.eu/Research/Library/ResearchDataServices/Guide (European University Institute Library 2021).

[9] See Wilkinson, Dumontier, Aalbersberg et al. 2016. Barend Mons provides a practical overview (Mons 2018).

JLIS.it

of variables, units of analysis, relationships between data elements, and whether or not the dataset is part of a series. Librarians also explain the importance of clear and consistent naming of folders, files, variables, versioning and documentation, and how good practice helps facilitate findability, accessibility, interoperability and reusability.

Supporting documentation should be updated throughout research projects because, when datasets are presented for reposit, documentation – such as codebooks and questionnaires – must also be submitted. Comprehensive documentation – describing dataset structure, folders, files, variables, versioning and (where applicable) information about problematic values, missing observations and weightings – makes research data findable, accessible, interoperable and re-usable (FAIR). Librarians – who are familiar with a wide variety of data documentation across disciplines and sub-disciplines – can offer feedback on data documentation and help edit dataset abstracts at the time of reposit.

Although all research institutions have data protection officers (DPOs), the library is frequently the first point of contact for scholars who have questions about database copyright and data protection. Librarians' long-standing experience with copyright and terms and conditions of access and use, has been extended to database copyright – which is important when library-licensed databases are used by researchers to generate new research data outputs. In many social science research projects, data outputs are the product of 'mixing' pre-existing data resources (frequently acquired and made available by the campus library) with new project-generated data (eg. surveys and experiments). For librarians, who are 'custodians' of subscription databases, it is important to inform database users of terms and conditions of access and use before research datasets based on licensed resources are openly shared.

During the research data lifecycle – and in collaboration with the DPO – librarians also inform scholars of their data protection obligations regarding the collection, use and security of data observations relating to persons, families and households. Librarians can advise on anonymisation and pseudonymisation techniques – which are particularly relevant for micro-level socio-economic data.

At the library of the European University Institute it is observed that many of the features of research data management (RDM) during the research project overlap with – and help prepare for – formal bibliographic control (BC) at the end – or near the end – of research projects.

4. Bibliographic control, infrastructure and workflow

Due to technological change, the exponential increase in digital content, and the momentum of the open access movement – universities began to establish institutional research repositories at the turn of the 21st Century.[10] Initially these infrastructures only indexed full-text documents and bibliographic records of publications. Gradually research dataset outputs and multi-media have been added, due to an ongoing research culture change towards open science and the increasing requirements of funding agencies. Research scholars also have the option to reposit

---

[10] Data on the growth of repositories (2000-2020) is available from the Registry of Open Access Repositories: https://en.wikipedia.org/wiki/Registry_of_Open_Access_Repositories. Accessed 7 April 2021.

JLIS.it

their data outputs in subject/domain repositories and 'catch-all' multi-disciplinary repositories, such as Zenodo.[11]

The EUI Library launched the Cadmus institutional repository, based on the DSpace infrastructure, in 2003.[12] The beta version of the EUI ResData repository was launched in 2016 and was merged with Cadmus in 2019. University librarians are increasingly aware of data-driven research projects because the campus library is the primary source for subscription databases; researchers usually require access to data software manuals provided by the library; data management plans are supported and reviewed by librarians and researchers frequently approach the library for advice on database copyright and data protection during the research data lifecycle.

However, it is not possible for librarians to be aware of every data-intensive project on campus, as there is no mandate for such information to be shared. Sometimes, librarians will only become aware of research data outputs when a principal investigator, or research team, approaches the library for advice on the preservation, reposit and open sharing of research dataset outputs – often due to funding agency requirements, such as the Horizon 2020 Framework Programme Open Research Data Pilot. Figure 2 provides an overview of the roles of researchers, librarians and ICT staff at the EUI during the research data lifecycle.

	ACTIVITY	RESEARCHERS	LIBRARY	ICT Service
Data input	Data discovery	Researchers discover data via library collections; the internet; and non-digital resources	Maintenance of data portal, indices and OPAC records	/
	Data generation	Researchers generate data (eg. surveys, experiments)	/	/
	Terms of access and use; database copyright and data protection	User compliance	Library promotes awareness of terms and conditions of access and use	ICT service and library provide access protocols
	Data management plans (DMPs)	Researchers write data management plans	Library provides training on DMP template tools and helps edit DMPs	ICT service provides standard description of infrastructure and security

---

[11] https://zenodo.org/. Accessed 7 April 2021.
[12] https://cadmus.eui.eu/. Accessed 7 April 2021.

JLIS.it

	ACTIVITY	**RESEARCHERS**	**LIBRARY**	**ICT Service**
Data elaboration / In-project data management	Dataset structure: folders, files, variables, observations	Researcher activity	Library advisory role	/
	Data anonymisation	Researcher activity	Library advisory role	
	Standardisation of file names, versioning, in-project metadata	Researcher activity	Library advisory role	/
	Documentation, codebooks and associated software/ routines	Researcher activity	Library advisory role	ICT advisory role
	In-project security and backup	Researcher activity	Library advisory role	ICT infrastructure and encryption software
Data output	Submitting research datasets	Researchers submit details of data outputs via online form	Library reviews submission	/
	Bibliographic control and metadata	/	Library checks structure and sources of dataset; converts submission information into metadata	/
	Repositing and infrastructure	/	Library reposits datasets in the institutional repository	Support for institutional repository infrastructure

Fig. 2. Research data lifecycle roles of researchers, librarians and ICT staff

4.1 Data submission to the institutional research repository

Over the past two decades there has been a growing awareness of the scientific value of making research data more openly available. While many academic disciplines have a long history of sharing underlying data within epistemic communities – the open science movement advocates wider access to research data as a public good, of benefit to scientific endeavour. Researchers in the social sciences and humanities are increasingly aware that the academic community is awarding recognition to research datasets as outputs in their own right. Reposited datasets can promote awareness of related publications, or in themselves become part of promotion and tenure procedures. In some cases, researchers become aware of these issues (open data; open science) late in the research project – for example, when an academic colleague or a funding agency requests information about underlying data. Researchers at the European University Institute who submit datasets for reposit are required to complete the library's online data submission form.[13] It is important to distinguish between three

---

[13] https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow. Accessed 7 April 2021.

JLIS.it

types of data description activities. Firstly, researchers generate essential descriptors for their data (names of folders, files, tabs, variables &c.) during the research project. These descriptors do not always constitute formal 'metadata' in the sense of bibliographic control. Secondly, the observations entered by researchers in the EUI library's online data submission form constitute 'raw' information about a dataset, and are never ingested directly into the repository without review. Thirdly, librarians generate bibliographic-standard metadata for the research repository; to make research datasets findable, accessible, interoperable and reusable.

This case study suggests that the in-project research data management (RDM) function complements the end-of-project bibliographic control (BC) function. The creation of synergies between the two library functions helps contribute to overall scientific quality control.

4.2 Initial review

The EUI library's online data submission form captures information which is used for verification, provenance and bibliographic control. At the EUI, the name and institutional email address of the principal investigator (or delegated submitter) is required for verification that the submitter is a member of the institution. Only works generated by EUI members – or research teams with at least one EUI member – can be included in the institutional repository. Alumni and former professors can submit datasets if the substantive part of the research was conducted while a full member of the university.

The EUI library became a member of ORCID, the Open Researcher and Contributor ID service, in November 2017. ORCID is a solution for authority control of authors' name variations across the EUI's Central Person Registry (CPR); the research repository Cadmus, and the ORCID registry. Both publications and datasets are associated with authors' ORCID IDs, providing increased visibly for researchers and the institution in the digital environment. EUI authors' names in the Cadmus repository are linked to the ORCID record – pushing publication and dataset metadata to their ORCID profiles.[14]

- When completing the dataset submission form, the names of all creators of the dataset must be listed – including technical collaborators if they have significantly contributed to the creation of the dataset.
- The title of the dataset submitted should not be identical to the title of a project or a publication. Librarians frequently offer suggestions regarding title clarity and, in many cases, titles are modified.
- The online submission form captures both the year of completion of the dataset (which may, or may not, be the current year); the date-range of data coverage – which is of great importance for any time-series data; and (where applicable) the geographical coverage of the dataset.
- Submitters are required to provide a description of the dataset – which is a first draft of the abstract displayed prominently in the repository entry.
- One of the most important submission form fields is the 'Source(s) of data'. If the dataset is the output of original data collection and elaboration, details should be provided. If the

---

[14] The complete workflow is explained by Lotta Svantesson and Monica Steletti, in their presentation at the Open Repositories conference (Svantesson and Steletti 2019). The EUI ORCID connect page is at: https://cadmus.eui.eu/ORCID/. Accessed 7 April 2021.

JLIS.it

dataset is derived from pre-existing sources, those sources should be clearly indicated (data creator, institutional source, publisher).

- The online submission form requires a preliminary statement of whether the data can be made available for open sharing immediately, or is to be reposited under embargo. Librarians help determine this status in consultation with researchers.
- Submitters are required to provide the file format of data files. If the data is in a proprietorial format, librarians can recommend (where possible) options for open format versions. This information is always translated into the related media type.[15]
- The number of data files within the dataset is an important field which allows librarians to discuss the relationship between the repository entry and the constituent elements. In some cases, it is necessary to create two entries (works) for a dataset which the submitter might be submitting as a single work. In other cases, multiple data submissions from the same project might be consolidated into one entry with multiple sub-sets.
- The online form also requests information regarding projected future waves of the dataset being submitted. This can require that a data sub-set which is intended to have future iterations, might need a separate entry in the repository.
- The online form also gathers information about supporting documentation, codebooks and (where applicable) software routines to enable the use of the data by others.
- The library advises on the appropriate reuse licence for open research data; eg: Creative Commons Attribution (CC-BY) or Public Domain (CC0).
- Submitters are asked to include references to related publications. This information can also be added when publications become available.

4.3 Provenance

The establishment of research repositories in universities and other institutions requires librarians to have a strong role regarding provenance. While research documents (working papers, theses, articles, chapters, monographs &c.) are normally subject to editorial review either inside the university or by external peer reviewers and publishers – the situation regarding research data outputs is more complex.

Very few universities have formal faculty-level 'editorial' review procedures for dataset outputs. The research data lifecycle is predominantly undertaken by researchers – with support from library and ICT professionals. At the end of research projects, the library becomes involved in issues of provenance, originality, data protection and database copyright. Here it can be seen that there is an overlap between the research data management (RDM) function and the bibliographic control (BC) function.

Although there are multiple ways in which librarians can undertake verification and provenance, it is impossible for librarians and information specialists to have detailed knowledge of data in every discipline and sub-discipline. It is also impossible for librarians to guarantee that every element and observation in a dataset is correct.

---

[15] Formerly known as MIME Type: https://en.wikipedia.org/wiki/Media_type. Accessed 7 April 2021.

JLIS.it

At the EUI, librarians build trust with researchers during the data lifecycle as part of the research data management function – informing scholars that;

> By submitting this [online submission] form, EUI members acknowledge that the dataset for reposit is the output of original data collection and elaboration; or is the output of significant, value-added, elaboration of pre-existing sources; and conforms with the EUI *Guide to Good Data Protection Practice in Research*.[16]

Librarians build trust with researchers through outreach and training; assistance with data management plans; provision of in-project services during the research data lifecycle and advice about database copyright, data protection, research ethics, scholarly reputation and scientific impact. When data is presented to the library for reposit, the 'Source(s) of data' field in the online submission form reveals whether the dataset output is partially based on pre-existing, library-licensed resources. At this point, it may be necessary for library staff to liaise with data suppliers to control for potential license issues regarding the open sharing of derivative datasets via the university repository.

4.4 Metadata generation

The generation of metadata about research datasets renders research datasets findable, accessible, interoperable and reusable (FAIR) and helps librarians decide whether research data outputs can be shared as open data. The research data management activities undertaken by librarians during research projects constitute a solid foundation for library bibliographic control and metadata generation. At the EUI, librarians use the raw information from the online dataset submission form to generate repository metadata using:

- The Dublin Core schema
- Library of Congress subject headings
- Dewey Decimal 23 classification
- A modified UN/Eurostat classification originally developed for the paper-format statistics collection[17] and,
- An internal data series identifier.

When setting up the EUI Research Data Collection structure in the Cadmus institutional repository, the EUI's institutional setup was reflected in the sequential, internal ID (dc.identifier.other) – eg: EUI_ResData_00032_HEC. The numeric value is a running sequence, with alpha-suffixes for:

- Economics: ECO
- History and civilisation: HEC
- Law: LAW
- Social and political sciences: SPS and
- The inter-disciplinary Robert Schuman Centre for Advanced Studies: RSC.

Here follows an example of the metadata record for a dataset reposited in 2020.

---

[16] https://www.eui.eu/documents/servicesadmin/deanofstudies/researchethics/guide-data-protection-research.pdf (European University Institute 2019).

[17] https://www.eui.eu/Research/Library/ResearchGuides/Economics/StatisticsClassification. Accessed 7 April 2021.

JLIS.it

Informal politics of codecision dataset	
dc.contributor.author	BRESSANELLI, Edoardo
dc.contributor.author	HERITIER, Adrienne
dc.contributor.author	KOOP, Christel
dc.contributor.author	REH, Christine
dc.coverage.spatial	European Union
dc.coverage.temporal	1999-2009
dc.date.accessioned	2020-09-09
dc.date.available	2020-09-09
dc.date.created	2014
dc.date.issued	2020
dc.identifier.other	EUI_ResData_00028_RSC
dc.identifier.uri	https://hdl.handle.net/1814/68095
dc.description	1 data file; 1 documentation file
dc.description.abstract	This dataset, created as part of the research project on 'The Informal Politics of Codecision' - funded by the Research Council of the European University Institute (EUI) and the Economic and Social Research Council (ESRC; Grant RES-000-22-3661) - is constituted by all 797 legislative files concluded under codecision between 1999 and 2009. It presents a new variable, 'early agreement', indicating whether legislation has been agreed informally, in trilogues, by the Council of Ministers and the European Parliament. It also includes variables with characteristics of the legislative file (legal nature, policy area, complexity, media salience, policy type, duration) and of the legislative negotiators (priorities of the Council Presidency, ideological distance between the Parliament's rapporteur and the national minister, the Presidency's workload).
dc.format	Excel file
dc.format.mimetype	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
dc.language.iso	en
dc.publisher	European University Institute, RSCAS
dc.relation.ispartofseries	EUI Research Data
dc.relation.ispartofseries	2020
dc.relation.ispartofseries	Robert Schuman Centre for Advanced Studies
dc.rights	info:eu-repo/semantics/openAccess
dc.rights.uri	http://creativecommons.org/licenses/by/4.0/
dc.subject	Legislative bodies
dc.subject.classification	FS-CA
dc.subject.ddc	328.4077
dc.subject.lcsh	Legislative bodies - European Union countries

JLIS.it

dc.title	Informal politics of codecision dataset
dc.type	Dataset
eui.subscribe.skip	TRUE
dc.rights.license	Creative Commons Attribution 4.0 International
dc.description.version	The dataset documentation is available in: BRESSANELLI, Edoardo, HERITIER, Adrienne, KOOP, Christel, REH, Christine, The informal politics of codecision : introducing a new data set on early agreements in the European Union, EUI RSCAS, 2014/64, EUDO - European Union Democracy Observatory -- Retrieved from Cadmus, European University Institute Research Repository, at: http://hdl.handle.net/1814/31612

Fig. 3. Example of metadata record, full view, from the EUI Cadmus repository Research Data Collection

This case study suggests that the in-project research data management (RDM) function complements the end-of-project bibliographic control (BC) function. The generation of metadata about research datasets helps to make research datasets findable, accessible, interoperable and reusable (FAIR). For example, the unique and persistent identifier helps researchers to find the dataset; the retrievability of the metadata via the repository protocol helps make the dataset accessible; the Dublin Core schema allows for broad sharing and interoperability, and the license information facilitates reusability.

4.5 Transfer and uploading of datasets

When EUI librarians have prepared the metadata record for the dataset, an appointment is made for the transfer of data, documentation and (where applicable) codebooks. At this stage, there may be further discussions about structure, format, provenance, copyright and data protection. Once the dataset is received and approved, the metadata file, the dataset and the documentation are uploaded in the Research Data Collection of the EUI Cadmus repository. A digital object identifier is generated and a data citation can be exported, eg:

> BRESSANELLI, Edoardo, HERITIER, Adrienne, KOOP, Christel, REH, Christine, *Informal politics of codecision dataset*, EUI Research Data, 2020, Robert Schuman Centre for Advanced Studies. Retrieved from Cadmus, European University Institute Research Repository, at: https://hdl.handle.net/1814/68095

The repository metadata schema allows further discovery of the resource, for example via library discovery tools and online aggregator services.[18] Accurate bibliographic control will also facilitate forthcoming machine discoverability of datasets and artificial intelligence applications.

---

[18] The EUI's Cadmus repository is interoperable with, and harvested by, CORE, Google Scholar, OpenAIRE, RePEC and Worldcat.

JLIS.it

5. Conclusion

Contemporary research data management is underpinned by the FAIR Guiding Principles, to make data findable, accessible, interoperable and reusable. Both the research data management (RDM) function and the bibliographic control (BC) function can be combined in the service of these principles.

Based on the experience of EUI library staff – this paper suggests that research data management during research projects and bibliographic control at the end of research projects are complementary elements of an emerging 'continuum' of library support for modern scientific culture – contributing to overall scientific quality control.

JLIS.it

References

Consortium of European Social Science Data Archives. n.d. "Data Management Expert Guide." Accessed 7 April 2021. https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide.

Digital Curation Centre. n.d. "DMPonline data management tool." Accessed 7 April 2021. https://dmponline.dcc.ac.uk/.

European University Institute. 2019. *Guide to Good Data Protection Practice in Research.* Accessed 7 April 2021. https://www.eui.eu/documents/servicesadmin/deanofstudies/researchethics/guide-data-protection-research.pdf.

European University Institute Library. 2021. *Research Data Guide.* Accessed 7 April 2021. https://www.eui.eu/Research/Library/ResearchDataServices/Guide.

International Society for Knowledge Organization. n.d. "Data". In *Encyclopaedia of Knowledge Organization.* Accessed 7 April 2021. https://www.isko.org/cyclo/data.

Joudrey, Daniel N., Arlene G. Taylor, and David P. Miller. 2015. *Introduction to Cataloging and mClassification.* 11th ed. Santa Barbara, CA: Libraries Unlimited.

Kellam, Lynda, and Kristi Thompson, eds. 2016. "Databrarianship: the Academic Data Librarian" In *Theory and Practice.* Chicago, IL: Association of College and Research Libraries.

Kruse, Filip, and Jesper Boserup Thestrup, eds. 2018. *Research Data Management: a European Perspective.* Berlin: De Gruyter Saur.

Mons, Barend. 2018. *Data Stewardship for Open Science: implementing FAIR Principles.* Boca Raton, FL: CRC Press.

Morriello, Rossana. 2020. "Birth and Development of Data Librarianship." *JLIS.it* 11 (3): 1-15. http://dx.doi.org/10.4403/jlis.it-12653.

OpenAIRE. n.d. "Argos data management tool." Accessed 7 April 2021. https://argos.openaire.eu/splash/.

Pradhan, Sanghamitra. 2018. *Cataloguing of Non-Print Resources: a Practical Manual.* New Delhi: Ess Publications.

Rice, Robin, and John Southall. 2016. *The Data Librarian's Handbook.* London: Facet Publishing.

Svantesson, Lotta, and Monica Steletti. 2019. "DSpace ORCID integration: name authority control solution at the European University Institute." Presented at the The 14th International Conference on Open Repositories (OR2019), Hamburg, Germany. https://doi.org/10.5281/ZENODO.3553926.

University of Southampton. "Registry of Open Access Repositories." Accessed 7 April 2021. http://roar.eprints.org/.

Wilkinson, Mark D., Michel Dumontier, Barend Mons et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Sci Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18.

JLIS.it

Integrated Search System:
evolving the authority files

Elena Ravelli[a], Maria Cristina Mataloni[b]

a) Istituto Centrale per il Catalogo Unico - ICCU, http://orcid.org/0000-0001-6402-2039
b) Istituto Centrale per il Catalogo Unico - ICCU, http://orcid.org/0000-0002-2791-2822

_____

Contact: Elena Ravelli, elena.ravelli@beniculturali.it;
Maria Cristina Mataloni, mariacristina.mataloni@beniculturali.it
Received: 1 April 2021; **Accepted:** 21 May 2021; **First Published:** 15 January 2022

_____

ABSTRACT

The coexistence of separate authority files within the main databases managed by the ICCU for entities of the same kind is going to be superseded in the new SRI portal through the integration, at the level of cooperative application, of the authority files for EDIT16 and Manus OnLine with those of SBN. The clustering of authority files, made possible through batch procedures and services provided by the applicative protocol SBNMARC, is intended to the development of browsable links between different representations of the same entity. The presence of identifiers and link keys between informative objects is therefore crucial to match data from the specialised databases EDIT16 and Manus OnLine, stored in the digital aggregator Internet Culturale, and shared through the collective catalogue SBN, with diverse quality and model but referred to the same resources and entities. The cluster of entities will be built upon the SBN Index, according to the quantity of data already available in its authority file and to exploit existing services and infrastructures which make shared cataloguing possible. SBN will also provide the spine of the integrated representation of entities through the public access platform of the new portal.

JLIS.it

After the 1966 flood in Florence, which caused extensive damage to the library collections and catalogues of the Biblioteca Nazionale Centrale di Firenze, the Centro nazionale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche, known since 1975 as the Istituto per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU), decided to proceed with the microfilming of card and paper catalogues in State libraries. This was the first step towards the digitisation of historical catalogues of the library collections held at preservation libraries; digitisation would guarantee the very survival of the catalogues and furthermore would offer access to their contents to a very wide public, even from a remote location.

The initiative aimed at averting the risk of dispersion of the immense cultural heritage available in Italian libraries. This task, in which ICCU is still engaged today, is very demanding if we consider that the Anagrafe delle Biblioteche Italiane[1] registers approximately 12,000 libraries of different types, divided among state and local authorities libraries, university libraries, ecclesiastical libraries and cultural institution libraries. This is a snapshot, not yet exhaustive, of the Italian library reality, characterised by an extreme fragmentation from a geographical, organisational and institutional point of view, which is the manifestation of the historical and cultural events of the country.

This is the context in which the SBN (Servizio Bibliotecario Nazionaled)[2], the network of Italian libraries, promoted by the Ministero dei beni e delle attività culturali e per il turismo in collaboration with regions and the university system, and coordinated by the ICCU, was born. It is based on an organisational model of cooperation and participation, designed to manage thousands of institutions. With this purpose in mind, in addition to the design and management of SBN and its database, the projects that led to the creation of ICCU's specialized bibliographic databases (EDIT16[3] and Manus Online[4]) started in the 1980s. All these projects have been crucial in bringing libraries out of their isolation and pushing them to build a network to obtain a mutual advantage in terms of visibility. At the same time, moreover, the cooperative model has proved to be the only viable way to guarantee the technological infrastructures, the expertise and financial resources capable of supporting the profound changes and the great complexity that have affected the field of Library and Information Science over the last decades.

Now ICCU is called to take a further step forward, that is, to enhance and make available the enormous work carried out over the years by Italian libraries to the widest and most heterogeneous audience possible, with particular regard to digital resources that are becoming increasingly important in number and quality. This is how the project Integrated Research System (SRI) was born, which foresees the possibility of querying ICCU databases at the same time through the creation of a single access point for searching and returning results. A fundamental aspect in this regard is the integration project of the database authority files managed by the Institute.

[1] Anagrafe delle Biblioteche Italiane. Accessed June 3, 2021. https://anagrafe.iccu.sbn.it/it/
[2] Servizio Bibliotecario Nazionale - SBN. Accessed June 3, 2021. https://www.iccu.sbn.it/it/SBN/
[3] Censimento delle edizioni italiane del XVI secolo - EDIT16. Accessed June 3, 2021. http://edit16.iccu.sbn.it/web_iccu/ihome.htm
[4] Censimento dei manoscritti delle biblioteche italiane. Accessed June 3, 2021. https://manus.iccu.sbn.it/

JLIS.it

Bibliographic databases of ICCU

Servizio Bibliotecario Nazionale – SBN

This infrastructural network is based on a stellar architecture whose centre is the Index (Indice), to which are connected the peripheral SBN Nodes (Poli), which include aggregations of libraries sharing resources, services, a user base and guidelines. The SBN Index, together with the management procedures, offers the services needed to establish the collaborative system that allows peripheral clients (the SBN Nodes) to share bibliographic information.

Currently there are nearly 6,600 libraries that have joined SBN, brought together in 104 Nodes. The size of the collective catalogue exceeds 18 million titles related to different types of material for a total of more than 102 million holdings.

Participation and cooperation are based on the sharing of common working methodologies, standards and uniform cataloguing rules as well as on a context characterised by flexibility in network participation. This flexibility allows Nodes to choose how to join on the basis of different profiles regarding the data to be shared with the collective catalogue. This context allows libraries to choose the quantity and quality of documents shared with the collective catalogue and the management of authority entries, and is essential in order to adhere to rules that make cooperation possible in a non-conflicting manner and consistent with the principles of the cataloguing involved.

On the catalogue front, ICCU has worked in recent years on the evolution of the SBN Index with the purpose of expanding the types of resources owned by libraries, including materials such as cartography, graphics, audio-visual, electronic resources and music. Moreover, in a highly modified context, both for the evolution of international standards and for the publication of the national cataloguing rules REICAT[5], the realisation of the authority file has received particular care and attention. As of 2018, specific regulations have been developed for the registration of the different entities at the authority level in SBN. The quality of SBN catalogue data is a key value for user services. With this in mind, ICCU is working to strengthen and differentiate forms of cooperation through specific working groups, coordinated by ICCU, which involve librarians from different institutions, representative of the entire national territory, in the review and implementation of SBN authorities[6].

EDIT16

The Censimento per le Edizioni Italiane del XVI secolo – EDIT16 was born in the 1980s with the purpose of documenting printed production from 1501 to 1600 in Italy and in the Italian language in other countries. Animated by a strong cooperative spirit, the project currently involves 1,597

---

[5] *Regole italiane di catalogazione REICAT.* 2009. Roma: ICCU. https://norme.iccu.sbn.it/index.php?title=Reicat
[6] The following working groups were launched in 2020: Working group for the management and maintenance of SBN Authority File. Printers, publishers, etc. (https://www.iccu.sbn.it/it/attivita-servizi/gruppi-di-lavoro-e-commissioni/gruppo-di-lavoro-per-la-gestione-e-manutenzione-dellauthority-file-di-sbn.-editori-tipografi-etc./) and Working Group for the management and implementation of the SBN Authority File. Names (https://www.iccu.sbn.it/it/attivita-servizi/gruppi-di-lavoro-e-commissioni/gruppo-di-lavoro-per-la-gestione-e-manutenzione-dellauthority-file-di-sbn.-nomi/)

JLIS.it

Italian and extraterritorial libraries. Since 2017, EDIT16 has expanded to record editions and holdings beyond national borders, confirming the transformation of the Census from a catalogue to a fundamental bibliographic resource for the study of Italian Renaissance culture. The first foreign institution to join EDIT16 was the British Library, which made a significant contribution to the database, thanks to the wealth of its collections of Italian editions.

In addition to the database reserved for bibliographic descriptions, the authority file, characterised by a highly specialised level of detail, has been developed including personal names, printing places, publishers/printers, and printer devices. EDIT16's authority file comprises structured data according to the descriptive standards of authority records with appropriately diversified elements to ensure their peculiarities. The quality and homogeneity of the data and the consistency of the information in the various databases have been ensured by the research work carried out by the editorial group, set up within the ICCU's Area di attività per la bibliografia, la catalogazione e il censimento del libro antico; since the beginning of the project, the group has taken on the management of the centralised collection of bibliographic descriptions, the definition of a working methodology, and the registration of authority records.

Two independent collateral databases, still closely related to the larger database, are the bibliography database, which describes the printed and electronic references listed in EDIT16, and the dedications database.

Digitisation is a crucial feature in EDIT16. More than half of the bibliographic records are complemented by title page and colophon images, and the description of each printer device includes its image. In recent years a considerable number of links to complete Italian and foreign digital copies available on the web has been added to EDIT16.

MANUS Online

Launched in 1988 with the coordination of the ICCU's Area di attività per la bibliografia, la catalogazione e il censimento dei manoscritti, Manus Online (MOL) is the first national project focused on the recognition and cataloguing of the immense manuscript heritage in the Latin alphabet from the Middle Ages to the contemporary age and preserved in Italian libraries. In recent years more and more space has been allocated for the census of papers (15th-20th centuries).

Since 2007, Manus Online has been available as an application arranged into specialised fields that include both a catalogue available to users and a cataloguing module, available free of charge for conservation organisations (public, private, ecclesiastical) participating in the project. Manus Online was the result of a collaboration among librarians, manuscript scholars and computer scientists who worked together to create a platform, which, all the while ensuring respect for the traditional descriptive elements of manuscripts, was more in line with international standards to allow an exchange of data and a dialogue with other bibliographic databases. Manus Online is based on an XML/TEI data schema, which was the most suitable format for exhaustively encoding the descriptive data of a manuscript and for allowing data to be exchanged with other data models as well, without loss of information.

There are 415 conservation and research organisations currently involved in the Manus Online project. In addition to librarians, individual scholars are also invited to propose variations to de-

JLIS.it

scriptive data through the Forum, a section of the portal that allows for a constant exchange of opinions and suggestions with ICCU and libraries.

Of particular interest is the section "Special projects": a module that allows the acquisition and management of specialist and international research projects. These projects, which maintain full organisational autonomy, use Manus Online as cataloguing software. The descriptions are offered to the public on each project's own platform as well as on the Manus Online website in the section reserved for such projects.

In 2015 the Gruppo di lavoro per la gestione e la manutenzione dell'authority file of Manus On-line was established with the purpose not only to monitor and correct the authority file in Manus Online but also to draw up guidelines with methodological procedures for the registration of authority records. The Manus Online section reserved for authority work is aimed at managing personal, collective, family and place names in printed sources and catalogued manuscripts.

Critical issues

As expressions of different projects, ICCU databases and information systems have been developed and managed over the years by separate offices. In the absence of central coordination to ensure an integrated development, these projects have been carried out using different softwares, languages and developmental approaches, thereby producing platforms based on very different data models and also data quality that meet the different expectations of their intended user communities.

Reference standards, even when held in common, have sometimes been adapted to the need to ensure the specificity of different objectives; moreover, in recent years, not all platforms have updated to the latest standards. As a result, at present, the same resource or entity can be recorded in different forms on individual platforms. A very clear example is the comparison between EDIT16 and SBN: as many users will have had the opportunity to verify, there is not always a biunivocal correspondence between the same resource described in the two databases. Just think of the case of different issues by date: in EDIT16 they are recorded in separate records whereas in SBN they have been described for many years[7] as variants within the same record. On the contrary, the same EDIT16 identifier can be associated with several SBN identifiers if in SBN a record has been created for each volume in a multi-level description, as well as for the main record, whereas only one record describing the whole edition has been created in EDIT16.

Even in cases where the name is recorded in the same form in the databases, their data model changes considerably. If we analyse, for example, the authority record of a printer including the same basic information, the structure of the data varies considerably. In EDIT16, in fact, the printers authority file, particularly significant for this resource, is organised into several fields and links, some of which are in SBN.

In order to access the resources of the individual ICCU databases, which are described separately in the different environments, the user needs to start from the specific search interfaces available,

---

[7] The more recent 2016 guidelines for early printed books in SBN, which have amended previous regulations, require that variants by date of the same edition should be described in different records. Still many corrections remain to be made on previous records.

in the absence of any connection among the platforms. The only exception, since 2016, is the introduction of the link between the bibliographic records of SBN and the corresponding records in specialised databases related to early printed books, including EDIT16. However, this link is not always available, being the result of individual cataloguing work.

Moreover, the specialised databases EDIT16 and Manus Online, and the SBN system allow for the management of digital attachments to descriptive records, but the current system of indexing and using digital resources (Internet Culturale and its index based on the BIB-MAG profile) does not recognise them. The latter makes integrated management of digital resources inefficient and complicated.

Internet Culturale: a partial solution

An integral part of the BDI (Italian Digital Library) project is *Internet Culturale, cataloghi e collezioni digitali delle biblioteche italiane*[8], a portal launched by ICCU in 2005 to promote knowledge of the Italian book heritage through access to both catalogues and digital collections. It also offers cultural insights through multimedia resources (itineraries, exhibitions, authors and works, 3D-paths), dedicated to literary, scientific, artistic and musical culture.

What you get from a search in Internet Culturale, in addition to the Digital Index, is a set of results returned from each database queried: EDIT16, Manus Online, SBN, Historical Catalogues[9] and Digital Library, the digital repository made available by ICCU for the Italian libraries. This is achieved by mapping data from ICCU databases on a common minimum Dublin Core profile. Search options are therefore limited, compared to what is offered by the different front-ends.

In view of the logical separation between records managed by different platforms, however, the Meta-Index system queries these databases, limiting itself to juxtaposing the information objects coming from the autonomous management environments. In addition, there are no engineered import procedures that ensure synchronisation in the alignment of Internet Culturale Indexes with those of SBN and of the specialised databases, thereby generating different search results by accessing the specific search sites.

Digital collections make up the Digital library, with more than 15 million associated digital files. The search in these databases is done through the Meta-Index of Internet Culturale by extracting data from the original databases. The data, whose characteristics, content and format vary, have been partially made consistent through a common profile based on the Dublin Core standard and qualified with the necessary extensions. However, the data profile in Internet Culturale is less rich than that of the same resources as they are recorded in SBN and the other specialist databases, and therefore does not fully represent and integrate the digital resources described through them. In addition, the data-feed process of the Digital Library of Internet Culturale is complex, rigid and rather expensive, especially for the institutions which do not have their own digital heritage management system and suppliers to provide this service.

---

[8] Internet Culturale. Accessed June 3, 2021. https://www.internetculturale.it/
[9] Cataloghi storici digitalizzati. Accessed June 3, 2021. http://cataloghistorici.bdi.sbn.it/

JLIS.it

Towards an integration of national bibliographic services: Integrated Research

In addition to the issues outlined above, there were cuts in professional resources, and, at the same time, the need to make new investments for the maintenance, updating and development of the individual platforms. As a result, we found ourselves having to rethink the structure of the information systems managed by ICCU in the twofold perspective of rationalisation and optimisation of resources, on the one hand, and of a new model of integration in data search and retrieval, on the other.

At the end of a process of reflection and in-depth analysis carried out within the framework of working groups promoted by the Direzione generale biblioteche e istituti culturali of the Ministero dei beni e delle attività culturali e per il turismo, the SRI project (Integrated Research System) was realised[10]. This project focuses on the integration of the databases managed by ICCU and at the same time will offer more effective services to meet the information needs of different user groups, ranging over from professional researchers up to the merely curious users. The project means therefore to overcome the fragmentation of the databases and rationalise the communication model of ICCU platforms, by creating a distributed information architecture that makes it possible to use the resources in the systems described above, including those in digital format, through a single access point.

The solution adopted to maintain information consistency among the elements of EDIT16, Manus Online and Internet Culturale Digital Index is achieved by configuring these systems in Client server mode with the SBN Index acting as a joint element and guide between these different systems. The new EDIT16 and Manus Online are, in fact, configured as specialist SBN nodes in the new set-up and are able to contribute to the collective catalogue while maintaining their own specificity and autonomy. The aim is to share a large part of the information with the reference record in the Index, with the future goal of sharing even greater information, once the system has been consolidated.

With this in mind, the back-end applications of EDIT16 and Manus Online have been re-engineered to enable dialogue with the central Index; at the same time, the Internet Culturale system has been optimised.

In order to achieve this goal, integration of the databases must be foreseen as early as in the creation phase of bibliographic records (only for EDIT16) and of the most important elements of the bibliographic data, such as names. The structural coherence between the databases is ensured by joint keys (mutual references) among the different information objects that refer to the same entities. This solution allows SRI to recognise that two representations refer to the same object. In particular, a resource X or an entity Y will be retrieved only once through the single search point, whereas specific representations, which are diverse for data models, for quality, or purpose, will be reached through links. Links will indeed allow navigation through the search interfaces, which will maintain their own functions and specificities designed for their own user community. As to the Manus Online database, the new re-engineered management application provides for the implementation of an exchange module exchange module, which allows sharing only authority

---

[10] For an overview of the steps that led to the development of the Integrated Research System, see Patrizia Martini. 2018. "Verso un'integrazione dei servizi bibliografici nazionali." *DigItalia* no. 2 (2018): 9-16.

records (including both intellectual responsibilities for texts and material responsibilities related to codicological features) with the SBN Index and which are aligned through specific services of the SBNMARC protocol, the dialogue protocol allowing the exchange of data between the central index and the SBN nodes. Any additional fields needed to complete the information in the Manus Online authority record are stored locally and not shared with the SBN Index.

Internet Culturale will continue to serve as an infrastructure dedicated to the collection and indexing of digital copies in its own repository and to those made available by remote repositories, but it will no longer have its own website, including a search engine and other tools available to users. In the architecture of the new information system, it will provide all those digital resources which can be linked to a cataloguing record in the main database to the central integration and indexing system, while also enriching the bibliographic core with all the technical and usage details in order to enhance its research tool on digital heritage.

Moreover, to widen the range of external providers of digital content, the aggregator will only allow the acquisition of descriptive metadata, leaving to the digital repository the function of making its digital content publicly available to users.

The process of authority files integration

As for bibliographic records coming from the SBN Index and the EDIT16 database, SRI has planned the unification of authority records. In particular, we refer to the files of personal and collective names of SBN, EDIT16 and Manus Online, and also, as far as the EDIT16 and SBN databases are concerned, to the printers devices and printing places.

Whereas inconsistencies between bibliographic records can be merely the result of cataloguing choices (e.g., the option of multilevel cataloguing), inconsistencies between names are rather substantial, and may prevent the ability of end users to identify an object to be the same in SBN, EDIT16 and Manus Online.

For instance, the authority records for Saint Roberto Bellarmino occurs in the three databases in three different forms:

1. Bellarmino, Roberto in SBN
2. Roberto : Bellarmino<santo> in EDIT16
3. Bellarmino, Roberto <santo ; 1542-1621> in Manus Online

The crucial role of a reconciliation among the three authority files is therefore clear, not only for end users but for cataloguers as well, who will be able to cross-reference authority records.

This process of integration of the authority files will be performed through a manual and painstaking double check of the databases. Some automatic procedures specifically developed will be useful at this stage to highlight problematic instances which require human intervention.

These procedures, which will concern the EDIT16, Manus Online and SBN authority files, will be addressed to catch duplicates or clearly erroneous forms. Such issues are rather frequent in SBN and Manus Online, whose data are the result of shared cataloguing, more exposed according to its nature to the risk of inconsistencies. This is not much of an issue for EDIT16, as its authority files are the result of research work carried out by the ICCU at a central level. However, EDIT16 records are often inconsistent with those of SBN from a formal point of view; indeed, EDIT16

JLIS.it

authority records have been created according to the previous Italian cataloguing rules for authors (RICA[11]) standards, as the database has been developed before the issue of the REICAT code in 2009. There has been no chance to update the authority records to the new cataloguing rules to date.

After these preparatory cleaning tasks, we will go ahead with the identification of SBN records matching records in specialised databases. In this process, SBN IDs will be added to the link field of the specialised databases, as well as the identifiers of the specialised databases to the SBN database. This will ensure data persistency, and their mutual reference will also be guaranteed. The inclusion of these relationships within the UNIMARC exchange files, meant to feed the new search interfaces, will make possible the aggregation of clusters of information objects by SRI's indexing engine.

The joint keys, which are built automatically within previous databases, will still be created when the new system is fully operational through the cooperation procedures. For this purpose, the matching algorithms, defined and refined within the development of the procedure described above and named 'Import-as-recognition', will also be included in the re-engineered specialised applications. The term 'import' means that names (of personal and collective entities) only included in EDIT16 and Manus Online will also be added to the Index database, which will store the whole set of authority records.

During the import procedure, the three authority files of SBN, Manus Online and EDIT16 will continue to be enriched and amended by the work carried out at the central level, as well as at a peripheral level by Nodes and libraries, both in the Index and in the specialised databases (as in cataloguing, data correction, etc.). In order to avoid the risk of misalignment and of jeopardising the consistency of the work done, the EDIT16 and Manus Online management systems will be re-engineered before performing the clustering procedures.

The reference information sheet, represented by the SBN record, is thus enriched with links to specialised databases, which, in turn, will have the chance to reach a larger audience, having their results listed not only in the new portal but also in the new OPAC SBN. Manus Online and EDIT16 will still maintain their own representation of entities in the reference systems as a result of the expertise of each database.

If these activities and tools prove to be apt to the task, similar procedures will be undertaken for other authority files, that is the printer devices database in EDIT16 and place names ones in both EDIT16 and Manus Online.

Integrated Research System – SBNTeca – Main services

As previously mentioned, the integration of the entire ecosystem also involves substantial correlation with digital resources and related metadata, stored and managed by the Digital Library system.

Therefore, an organisation of the complex system of aggregation and fruition of the digital resources of ICCU and Internet Culturale is in progress, as a companion to the development of the

---

[11] *Regole italiane di catalogazione per autore.* 1982. Roma: ICCU.

JLIS.it

new management applications. Basically, the new architecture of the digital flow allows digitised items to be in fact an extension of the catalogue through the link between digital copies and the bibliographic records, which was often missing in the information set provided in the previous systems.

Another fundamental component of the SRI is therefore the SBNTECA, which is a digital library capable of allowing the management of digital objects (images, audio-visual documents, etc.) within individual SBN Nodes and their exposure to the central SRI system and, from here, their display through the central SRI system as well.

SBNTECA, besides allowing the management of digital objects, is also used for the creation, import and management of metadata associated with such digital objects (technical metadata), as well as aggregates between them (e.g., the pages of one book). To do this, it must be able to act on the main metadata standards for digital content: MAG, a standard for management and administrative metadata, and METS (mainly in the Google-METS and METS-ICCU specifications).

The services made available by SBNTECA are also meant to recover the 'submerged digital', often included in often difficult to access share due to preservation contexts, as well as poorly valued or only available in off-line environments managed by Italian libraries. Also, the new digitisation campaigns find, in this context, an efficient system of management and fruition of digital and multimedia content.

Network of portals[12]

The multi-layer work described so far, mainly addressed to reconcile authority files, proves to be meaningful for end users in the results of the new platforms. Each database presents indeed a new faceted search interface providing features made possible by the new architecture. It is, in fact, a network of portals, each of which gives access to research services and types of content intended for different communities of users characterised by diversified information needs.

The case of Manus Online is significant under this respect, as it offers a more articulated internal representation of the manuscript record and a clearer identification of the textual units and their grouping in codicological units.

Functionalities will not be weakened; on the contrary, the system as a whole will provide new search options in a more technologically advanced and optimised context.

SBNTECA's services, which are integrated into the management systems of specialised databases, make direct representation of digital content available through the ecosystem's central viewer, Mirador, based on the IIIF protocol.

Alongside the re-engineering and functional review of the specialised research platforms, the project has planned a similar intervention on the portal of the SBN catalogue. The main difference is that the SBN OPAC will provide an integrated search within the bibliographic records of EDIT16 and Manus Online. The SBN catalogue will be complemented by links to records coming from EDIT16 and Manus Online; EDIT16 records can either be referred to resources also includ-

---

[12] Cerullo, Luigi, and Maria Cristina Mataloni. 2020. "Sistema di ricerca integrato: un nuovo catalogo di servizi per le biblioteche." *DigItalia* no. 2 (2020): 16-25.

JLIS.it

ed in SBN as part of the collections of one or more SBN libraries, or to resources not included in SBN when part of collections of libraries not participating in SBN. Manuscripts coming from Manus Online, on the other hand, are resources of a type not managed in the collective catalogue to date. In this case, internal descriptions of manuscripts are anyway retrieved, i.e. records including bibliographic elements and authorities, whose descriptive profile is better suited to the data return model of the shared catalogue.

ALPHABETICA: a new portal for Italian libraries[13]

To the eyes of end users, the most relevant new tool will be the new portal of Italian libraries, Alphabetica. The bibliographic core, represented by the general catalogue and its logic integration, is enriched by the data coming from other related databases managed by the ICCU, such as the portal *1418 – documenti e immagini della grande guerra14*, the historical catalogues, virtual exhibitions built with *Movio– Mostre virtuali Online15* (Digital Online exhibitions). Moreover, the architecture of Alphabetica will possibly allow the integration of more databases potentially interested in joining the Alphabetica network. The logic behind the portal goes beyond the traditional model of bibliographic research and restitution of an OPAC, even of an advanced one, and is an attempt to build a proper search model based on a solid but flexible system of taxonomies.

Alphabetica classifies all kind of resources and related entities (names of collective entities, places, persons) according to a dual classification system to comply both with the traditional classification of material in SBN, and with controlled vocabularies for the classification of objects in order to arrange them within thematic channels. This approach will provide a way round to the oddity of semantic terms in the catalogue, which we are in the meantime trying to address through rather complex off-line procedures on the SBN Index.

The stages required for the analysis, planning and testing of Alphabetica are obviously along and painstaking process in order to realise an innovative and effective new reference, which will be able to exploit and showcase the longstanding and valuable work carried on in Italian libraries, not only for a specialist audience but open to a new and diverse user base.

---

[13] Buttò, Simonetta. 2020. "Alphabetica, il nuovo portale per la ricerca integrata: un salto di qualità per le biblioteche italiane." *DigItalia* no. 2 (2020): 9-15.

[14] http://www.14-18.it/home

[15] https://www.movio.beniculturali.it/

JLIS.it

References

1418 Documenti e immagini della grande guerra. Accessed June 3, 2021. http://www.14-18.it/home.

Anagrafe delle Biblioteche Italiane. Accessed June 3, 2021. https://anagrafe.iccu.sbn.it/it/.

Buttò, Simonetta. 2020. "Alphabetica, il nuovo portale per la ricerca integrata: un salto di qualità per le biblioteche italiane." *DigItalia*, no. 2 (2020): 9-15. http://digitalia.sbn.it/article/view/2624

Cataloghi storici digitalizzati. Accessed June 3, 2021. http://cataloghistorici.bdi.sbn.it/.

Censimento delle edizioni italiane del XVI secolo - EDIT16. Accessed June 3, 2021. http://edit16.iccu.sbn.it/web_iccu/ihome.htm.

Censimento dei manoscritti delle biblioteche italiane. Accessed June 3, 2021. https://manus.iccu.sbn.it/.

Cerullo, Luigi, and Maria Cristina Mataloni. 2020. "Sistema di ricerca integrato: un nuovo catalogo di servizi per le biblioteche." *DigItalia*, no. 2 (2020): 16-25. http://digitalia.sbn.it/article/view/2625

Internet Culturale. Accessed June 3, 2021. https://www.internetculturale.it/.

Martini, Patrizia. 2018. "Verso un'integrazione dei servizi bibliografici nazionali." *DigItalia*, no. 2 (2018): 9-16. http://digitalia.sbn.it/article/view/2162.

MOVIO Mostre Virtuali Online. Accessed June 3, 2021. https://www.movio.beniculturali.it/.

Regole italiane di catalogazione per autore. 1982. Roma: ICCU.

Regole italiane di catalogazione REICAT. 2009. Roma: ICCU. https://norme.iccu.sbn.it/index.php?title=Reicat

Servizio Bibliotecario Nazionale - SBN. Accessed June 3, 2021. https://www.iccu.sbn.it/it/SBN/.

JLIS.it

DREAM. A project about non-Latin script data

Antonella Fallerini[(a)], Agnese Galeffi[(b)], Andrea Ribichini[(c)], Mario Santanché[(d)], Mattia Vallania[(e)]

a) Sapienza Università di Roma
b) Sapienza Università di Roma, https://orcid.org/0000-0003-0799-5699
c) Sapienza Università di Roma, https://orcid.org/0000-0002-0281-4257
d) Sapienza Università di Roma, https://orcid.org/0000-0003-1777-1162
e) Sapienza Università di Roma

_____

Contact: Antonella Fallerini, antonella.fallerini@uniroma1.it; Agnese Galeffi, agnese.galeffi@uniroma1.it; Andrea Ribichini, ribichini@diag.uniroma1.it; Mario Santanché, mario.santanche@uniroma1.it; Mattia Vallania, mattia.vallania@uniroma1.it
Received: 5 May 2021; **Accepted:** 16 June 2021; **First Published:** 15 January 2022

_____

ABSTRACT

The DREAM project is a large research project funded by Sapienza University of Rome, dealing with bibliographic data in non-Latin scripts. As the National Bibliographic Service catalogue (SBN) does not yet manage data in non-Latin scripts, the aim of DREAM is to offer researchers a catalogue searchable through original scripts (such as Arabic, Chinese, Cyrillic, etc.). One of the most remarkable features of the project is the creation of an ILS-independent working context in which the cataloguer may find and retrieve data in original script from authoritative catalogues, starting from the existing romanized ones. From a technical standpoint, the ever increasing Unicode support offered by modern operating systems, DBMSs and indexing engines makes the rapid development of the relevant software tools a concrete possibility. This in turn implies a shift in scientific focus towards the (often subtle) record linkage operations between different data sources. The authors hope that the DREAM project will gather the adhesion of other Italian libraries that perceive the same needs. Furthermore, as soon as SBN will support the management of data in non-Latin scripts, the DREAM project partners will be able to contribute with their data.

JLIS.it

Non-Latin script cataloguing. The context

The DREAM (Data Recording Entry Alternative Multi-script) project was born in Sapienza university in 2019 in order to create a repository for bibliographic data in non-Latin scripts, publicly available as a cooperative catalogue. This need arises from the evidence that the SBN national catalogue, to which Sapienza libraries adhere as do thousands of other Italian libraries, does not fully support the UTF-8 character encoding. SBN catalogue is based on shared cataloguing: all the participant libraries contribute sending data from the local nodes (that is, aggregation of libraries) to the central index. Member libraries may use a variety of LMS authorized by the ICCU, the national agency in charge with the SBN catalogue maintenance, but regardless of the software capabilities, the central index accepts data in Latin script only. All the languages expressed through other scripts, such as Arabic, Chinese, Cyrillic, Hebrew, Japanese, Greek, must be transliterated using the ISO instructions. This requirement is stated in the Italian cataloguing rules, REICAT (ICCU 2016b), and restated by the SBN cataloguing instructions (ICCU 2016a). The transliteration process has many disadvantages for both the involved actors – cataloguers and users.

Notwithstanding some attempts at automatic transliteration (Eryani 2021) and the availability of online tools (DuBose 2019), this activity is a very time consuming one presenting a large number of technical problems (Ismail and Md. Roni 2010), not to mention the variety of connected cataloguing issues such as

- The use, in some contexts, of unsound transliteration scheme (Molani 2006).
- The transliteration and conversion of personal names, place names, corporate bodies, and other entities (Li 2004).
- The subject access (El-Sherbini and Chen 2011).

Besides that, the equity of access – one of the basis of the ethics in library science – is not guaranteed since those who need these materials have to determine how cataloguers could have transliterated the data. On the opposite, the users who want roman script resources are not required to make this extra and inefficient effort (Agenbroad 2006, 22). For users who are native language speakers in non-Latin scripts, the transliteration is totally useless since they have all the knowledge and tools to perform a search in the library catalogues using the original script.

For all other users reading in second language, transliterated text requires an additional cognitive demand since they typically acquire and access a cohesive set of phonological, orthographic (and possibly semantic) representations of words in their second language, whereas transliteration requires readers to create cross-script associations between phonological-semantic representations in one language and previously unrelated orthographic forms in another (Rao, Mathur, and Singh 2013, 205). All these elements cause many obstacles in users' searching ability and accessibility in retrieving bibliographic records (Kim 2006)

The DREAM project

The DREAM project is a major university project funded by Sapienza University of Rome in 2019; the project leader is Federico Masini, professor of Chinese at the Istituto Italiano di studi orientali (ISO) of Sapienza university of Rome and the research staff is a mix of academics and library per-

JLIS.it

sonnel, both involved on the front line of the activities. The very first idea of the DREAM project arose in May 2018, when Antonella Fallerini, librarian at the ISO library, joined the workshop "Building a Network of Korean Resources Specialists in Europe", organized by Freie Universität Berlin – Campus Library and funded by the Korea Foundation. The workshop aimed at bringing together European Korean Studies librarians in order to develop a professional network within Europe and strengthen the representation of interests of Korean Studies librarians in national and worldwide library information structures. While discussing with the colleagues attending the workshop, Fallerini highlighted the severe limitations of transliterated data compared to bibliographic descriptions in original scripts. The expected result of that was that some colleagues confirmed they did never find a single record in original script in Italian catalogues. Their obvious self-explanation was that there was a great scarcity of our collections in Far Eastern languages. The available online catalogues do not give any advice about the transliteration and researchers have no reason to expect such a treatment. The extensive transliteration practice in cataloguing has as a direct consequence the underrepresentation of our library collections both at national and international level. To give an idea of the extend of the phenomenon, an internal preliminary investigation conducted on the online catalogues of the most representative Italian institutions, such as larger universities and research libraries, have shown that more than 500,000 resources have been catalogued in romanization. It is possible to estimate that there are at least double that number waiting to be catalogued.

What is the aim of DREAM?

DREAM project aims to figure out a provisional and cooperative solution in order to create a repository for non-Latin scripts data, available as a catalogue in the near future. At the present moment, the cataloguers must create transliterated data to feed SBN; if adhering to the DREAM project, the cataloguer will also use DREAM tools to search for the corresponding record in original script in authoritative catalogues. Once the possible matches have been identified, the system will present them to the cataloguer, giving him/her the responsibility of confirming of dismissing them. These data will complement the transliterated ones that are already being produced for SBN shared cataloguing.

The result of this procedure will be a cluster of records that will be show in the DREAM catalogue giving the user the possibility to make searches using the preferred forms (in original scripts or in transliteration, even according to different schemas).

The DREAM project do not want to propose an SBN-alternative context. On the opposite, its features are developed taking into consideration both the respect of SBN rules and its potential developments. When SBN will accept data in non-Latin scripts, the libraries adhering to DREAM will have the possibility to feed their records into the national catalogue. This is why the DREAM project has a provisional nature. The adjective "provisional" may be referred to two aspects: first of all, it connotes the research aspect. DREAM's aim is to produce a working solution and at the same time, to explore, to verify, and to find the best ways to achieve the projects' goals. Since Sapienza libraries are part of the SBN network, there is no intention to create a new network or some alternative solutions; this is the second significance of provisional. DREAM project wants to create an environment where the cataloguer can retrieve data in

JLIS.it

non-Latin scripts, match them with the available transliterated ones and make them available in a specific DREAM catalogue. Both the working environment and the user search interface will be independent from SBN as well as from the software used by the libraries joining the project. We hope in fact that other Italian libraries – even those not members of SBN – will be interested in joining the DREAM catalogue, once some of the fundamental components of the its architecture have been realized. The DREAM project is still ongoing. We would like to stress one of the project's strengths: flexibility. We have a clear idea of the final results we want to achieve, but there is no prejudice about how to reach them. There are just some constrains due to the cataloguing context we have to dialogue with at some point, that is the software used and the SBN catalogue.

Main points

DREAM is ILS independent

Commercial software available to librarians are built to maximise output (cataloguing, lending, library management, etc.) and are therefore, in most cases, designed to be stable and standard. If you need a flexible environment to, for instance, carry out an experiment or a research project, it is difficult to balance these development needs – maybe even unsuccessfully – with the commercial logic of software distributors. Anyway, Sapienza has invested in our ILS (SebinaNext) in order to implement in the near future some new features, such as to accept, manage and visualize data in non-Latin script and especially right-to-left scripts, to handle VIAF id and an OAI-PMH module for authority data. DREAM will be an external and ILS independent environment. This need would not have arisen if we had a flexible and welcoming open source software or library platform in use. In this case, the DREAM project would have been just another component, a small one, of a larger system. What we learnt: often the paths you thought you were taking do not turn out to be fruitful and you have to go back, change your path and sometimes even rewrite the map. These features of research projects do not match the market logic.

Retrieve bibliographic data from reliable sources

In order to quickly populate the DREAM catalogue, we are going to start from the traditional transliterated records already existing, to search for equivalent records in original script in authoritative catalogues, and to import them. These procedures present certain degrees of difficulty. First of all, the identification of reliable sources to retrieve data. This is a scientific but also a technical task. It is not only a matter of knowing the most representative institutions for the languages of interest, but also of selecting those that have a data format easy to manage or map and an accessible retrieval option.

The current DREAM implementation supports this "search and match" between, on one side Sapienza University of Rome catalogue and on the other side the Bibliothèque nationale de France, the Library Union Catalogue of Bavaria, Berlin and Brandenburg (B3kat), and the Système universitaire de documentation (SUDOC). Since these sources expose their data through a variety of protocols, such as OAI-PMH, SRU and Z39.50, different clients are needed. More-

JLIS.it

over, to process data, specific response parsers are required for each source. As a matter of fact, even though the retrieved data are in standard formats (MARC21, UNIMARC), the packaging of data varies from source to source, containing error messages and paging information in ad hoc formats. Even in environments that we assume to be highly standardized (dealing with MARC, Z39.50, SRU, OAI-PMH formats) we found, in addition to the expected MARC21-UNIMARC dichotomy, USMARC or local dialects of MARC, Dublin Core, and several application profiles. In order to obtain a presumed match of the data, different analyses and mappings are required each time for their retrieval and processing.

Different sources (we are talking about national bibliographies/catalogues and national library catalogues) also have different approaches to standards.

For example, MARC21 allows to put in the same record data in the original script (e.g. Cyrillic) together with transliterated data by using the combination 880 and $6 but the cataloguing agency can choose whether to put in 880 the original script or the transliterated version. This allows the creation of (at least) two versions of the record. Moreover, the different granularity of the data contributes to make the match uncertain.

Authority data

Obviously, within the DREAM environment, in addition to bibliographic data, it is essential to import, manage and use authority data. In this respect, VIAF is the point of reference. Since the VIAF id is widely used, it is not only possible to retrieve authority clusters, but also to use the VIAF id as a bridge to navigate through catalogues in search of other bibliographic data of potential interest.

What we are building. The DREAM architecture

We designed a flexible, modular and scalable software architecture for a multiscript MetaOPAC (see Figure 1), based on the data warehousing paradigm (Inmon 2005; Kimball et al. 2008). We also developed a prototype implementation for research purposes (i.e., feasibility assessment, experimental evaluation of adopted solutions). The following is the description of our architecture's main components.
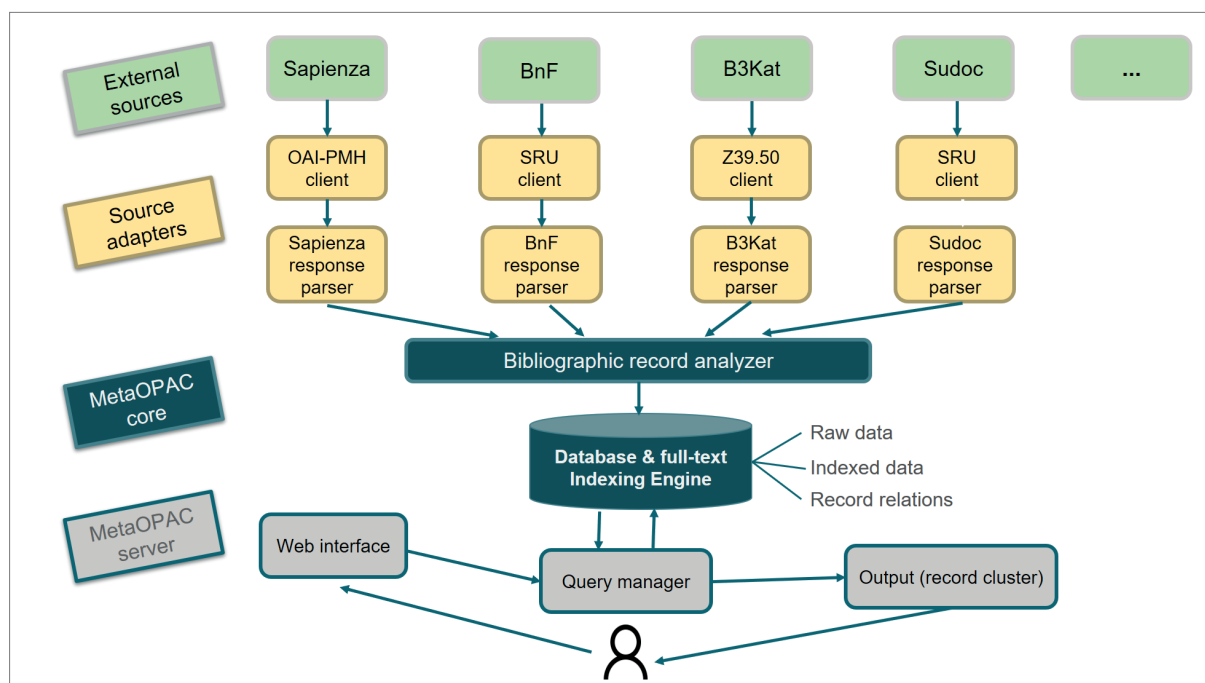
JLIS.it

Fig. 1. MetaOPAC architecture

Source Adapters. We have taken into consideration and tested several data sources. The current implementation supports Sapienza University of Rome's own catalogue, the Bibliothèque nationale de France (BNF), the Library Union Catalogue of Bavaria, Berlin and Brandenburg (B3Kat), and the Système universitaire de documentation (SUDOC). These sources make their data available through a variety of protocols, such as OAI-PMH, SRU and Z39.50. Therefore, clients are needed for each of these protocols. Moreover, an ad hoc response parser is required for each data source. This is because, even though the returned data are provided in standardized formats (e.g., UNIMARC, MARC 21), the packaging of these formats varies from source to source.

MetaOPAC Core. Downloaded bibliographic records are stored in a (relational) database, analyzed, and portions that are relevant for future search queries, e.g., title, authors, publisher (including all variants in both native script and transliteration, if present) are saved separately and properly indexed. Our prototype implementation currently uses MySQL as DBMS. The database structure consists of three tables:

- Table "raw" contains the unprocessed downloaded records.
- Table "indexed_data" contains, for each record, the extracted data to be indexed in order to speed up searches. At the present moment, we rely on MySQL's full-text indexing capabilities (a recent addition). We remark that different scripts require different indexing methods: alphabetic and syllabic scripts are handled by the default token-based full-text indexer, with minimum token size set to 1 and stop words exclusion disabled, while Ideographic scripts are instead dealt with by an n-gram based indexer, with n=2.
- The third database table, "relations" represents associations between records from different data sources, that we call "clusters". Clusters may be established through several methods (that we will discuss shortly).

JLIS.it

MetaOPAC Server. Searches in our prototypal MetaOPAC implementation can be run through a web server that accepts HTTP GET requests. In addition to the traditional search criteria (keywords, title, author, publisher), wildcards and boolean operators are accepted. A query manager translates the searches into full-text database queries. The search results are returned as an XML document listing retrieved clusters sorted by *relevance* (a measure of the adherence of the records in each cluster to the search criteria).

How to feed the DREAM. Record linkage among data sources

In our MetaOPAC application, the construction of clusters (i.e., groups of records referring to the same entity) may be carried out through three methods.

1. *Manual Intervention.* The cataloguer manually identifies the correspondences between records from different data sources. In our prototype we have created 27 Sapienza-BNF pairs, 27 Sapienza-B3KAT pairs, and 40 Sapienza-SUDOC pairs. It is hoped that, as the number of partners grows, more and more librarians will contribute their associations across data sources to the MetaOPAC database.

2. *Identification by Unique Identifiers.* A second way to identify correspondences between records from different data sources is through unique identifiers. In our prototype we have used ISBN to search for matches (all supported external catalogues allow ISBN-based searches through their APIs).

Document Language	Sapienza Records with ISBN	Sapienza-BNF ISBN-based Matches	Sapienza-B3Kat ISBN-based Matches	Sapienza-SUDOC ISBN-based Matches
ARA	369	122 (33.06%)	113 (30.62%)	126 (34.15%)
CHI	1875	98 (5.23%)	492 (26.24%)	399 (21.28%)
HIN	25	8 (32%)	3 (12%)	8 (32%)
JPN	1771	246 (13.89%)	692 (39.07%)	781 (44.10%)
KOR	2191	80 (3.65%)	432 (19.72%)	457 (20.86%)
PER	66	7 (10.61%)	12 (18.18%)	15 (22.73%)
SAN	73	17 (23.29%)	26 (35.62%)	28 (38.36%)
SWA	1	1 (100%)	1 (100%)	1 (100%)

Table 1. Breakdown of positive search results

3. *Algorithmic Techniques.* The third method consists of a blend of well-established record linkage algorithmic techniques and ad hoc solutions. We proposed the following workflow, based on the VIAF:
   - Given as input a bibliographic record, we extract the VIAF code of its author (assumed to be present).
   - We then run a search on the VIAF online service for the extracted id, obtaining the variant form of the author's name used by each data source.

JLIS.it

- For each supported source, a search-by-author, using the variant form obtained through VIAF as input string, is performed. This allows us to restrict the search domain to the works of that author.
- Finally, we run any record linkage algorithm we see fit in order to identify the correct matches between the input record and the records retrieved from the other data sources.

Standard record linkage techniques include the use of string similarity measures (Navarro, 2001) – Levenshtein distance (Levenshtein, 1966) being a popular one – to assess correspondences between fields such as title, subtitle and publisher (including their variants and versions in original script, if present). Comparison of other metadata (e.g., publication dates) may also be useful as a verification tool. Moreover, if the bibliographic record belongs to a cluster in the MetaOPAC database, then all metadata of the cluster may be used to identify the correct match. More sophisticated, domain-specific techniques may include transformations from one transliteration standard to another, and switching from original script to transliteration and vice versa. Early testing on 19 Sapienza records, manually matched with both BNF and SUDOC to provide a "ground truth", has shown correct results in 17 cases. This is quite promising considering that for this test only the minimum normalized Levenshtein distance (i.e., Levenshtein distance divided by the length of the longest input string) between all title variants has been considered as a criterion.

Further steps

The project next steps are:

- Engaging partner institutions: we hope that this conference will also be an opportunity to promote the project and involve other partners who share the problem with data in non-Latin scripts
- From a technical standpoint, further tasks would include writing adapters to support additional sources, and launching larger scale algorithmic record linkage runs with feedback loops involving manual sample validation and fine-tuning of algorithmic features. Identified clusters should then be fed into the MetaOPAC prototype implementations, with measurement of both load and query times, in order to determine performance-critical sections that may need refinement both at the implementational and the architectural level.
- It is also needed to develop all the interfaces, both the back office minimal interface to allow cataloguers to validate the matches between records and the public DREAM catalogue search interface.

JLIS.it

Bibliography

Agenbroad, James E. 2006. "Romanization Is Not Enough." *Cataloging & Classification Quarterly* 42 (2): 21-34. https://doi.org/10.1300/J104v42n02_03

DuBose, Joy. 2019. "Russian, Japanese, and Latin Oh My! Using Technology to Catalog Non-English Language Titles." *Cataloging & Classification Quarterly* 57 (7-8): 496-506. https://doi.org/10.1080/01639374.2019.1671929

El-Sherbini, Magda, and Sherab Chen. 2011. "An Assessment of the Need to Provide Non-Roman Subject Access to the Library Online Catalog." *Cataloging & Classification Quarterly* 49 (6): 457-483. https://doi.org/10.1080/01639374.2011.603108

Eryani, Fadhl, and Nizar Habash. 2021. "Automatic Romanization of Arabic Bibliographic Records." https://arxiv.org/pdf/2103.07199.pdf

ICCU. 2016a. "Guida alla catalogazione in SBN – Materiale moderno." Last modified July 13, 2016. https://norme.iccu.sbn.it/index.php?title=Guida_moderno/Descrizione/Capitolo_generale/Lingua_e_scrittura_della_descrizione

ICCU. 2016b. "Regole italiane di catalogazione. Appendice F – Traslitterazione o trascrizione di scritture diverse dall'alfabeto latino." Last modified September 21, 2016. https://norme.iccu.sbn.it/index.php?title=Reicat/Appendici/Appendice_F

Inmon, William H. 2005. *Building the data warehouse.* 4th ed. Indianapolis: John Wiley & Sons.

Ismail, Mohd Ikhwan, and Nurul Azurah Md. Roni. 2010. "Issues and challenges in cataloguing Arabic books in Malaysia academic libraries." *Education for Information* 28 (2-4): 151-163.

Kim, SungKyung. 2006. "Romanization in Cataloging of Korean Materials." *Cataloging & Classification Quarterly* 43 (2): 53-76. https://doi.org/10.1300/J104v43n02_05

Kimball, Ralph, Margy Ross, Warren Thornthwaite, Joy Mundy, and Bob Becker. 2008. *The data warehouse lifecycle toolkit.* 2° ed. Indianapolis: John Wiley & Sons.

Kudo, Yoko. 2010. "A Study of Romanization Practice for Japanese Language Titles in OCLC WorldCat Records." *Cataloging & Classification Quarterly* 48 (4): 279-302. https://doi.org/10.1080/01639370903338352

Levenshtein, Vladimir Iosifovich. 1966. "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady* 10 (8): 707-710.

Li, Yue. 2004. "Consistency versus Inconsistency: Issues in Chinese Cataloging in OCLC." *Cataloging & Classification Quarterly* 38 (2): 17-31. https://doi.org/10.1300/J104v38n02_04

Molavi, Fereshteh. 2006. "Main Issues in Cataloging Persian Language Materials in North America." *Cataloging & Classification Quarterly* 43 (2): 77-82. https://doi.org/10.1300/J104v43n02_06

Navarro, Gonzalo. 2001. "A guided tour to approximate string matching." *ACM Computing Surveys* 33 (1): 31-88. https://doi.org/10.1145/375360.375365

Rao, Chaitra, Avantika Mathur, and Nandini C. Singh. 2013. "'Cost in Transliteration': The neurocognitive processing of Romanized writing." *Brain and Language* 124 (3): 205-212. https://doi.org/10.1016/j.bandl.2012.12.004

JLIS.it

Two Projects and a Thesaurus.
Recent Experiences in the Management, Description and Indexing of Oral Sources

Sabina Magrini[(a)]

a) Ministero della Cultura

ABSTRACT

The Istituto Centrale per i Beni Sonori e Audiovisivi (ICBSA) has just finished, together with the Università degli Studi di Siena and Università degli Studi di Siena per stranieri, to work on the project "Ti racconto in italiano" which focuses on providing different access points to audio resources collected between 1980's and 2000's by ICBSA itself, as part of its mission to document Italian audio and audiovisual culture.

The main aim of the project is to create tools which will enable scholars of social history, art and literature to use these sources as well as providing original material for foreign students to exercise their knowledge of Italian. In order to facilitate access it has been necessary to create finding aids such as indexes and thesauri. For this purpose ICBSA has started a collaboration with the Biblioteca Nazionale Centrale di Firenze and the latter's Nuovo Soggettario.

This is not the first case of a project by insititutes of the Italian Ministry of Culture comprising the use of the Nuovo Soggettario for the indexing of archival materials. Indeed, the Soprintendenza Archivistica e Bibliografica della Toscana has already worked in this direction a few years ago when treating the so-called Straw archives.

KEYWORDS

Nuovo Soggettario; Oral sources; Archives; Indexing.

JLIS.it

The aim of this paper is to address a number of significant issues concerning the main theme of this Conference on bibliographic control in the digital ecosystem and mainly:

1. The complex interactions that are becoming common between different areas of knowledge and knowledge management in the bibliographic universe;
2. New ways of indexing documents;
3. The role of Thesauri in digital systems.

To do so, it shall be necessary to concentrate at first on the project "Archivi di paglia" which the Soprintendenza archivistica e bibliografica della Toscana developed around 2014-2016 in collaboration with the Biblioteca Nazionale Centrale of Florence. For those who may not be familiar with the intricacies of Italian cultural administration, the Soprintendenza is a Supervision Agency, the local office of the Italian Ministry for Culture engaged in the protection and valorisation of notified archives and libraries belonging to private individuals or archives and libraries belonging to public (non-State) entities in Tuscany. Object of this project was the census of the companies (and their archives) which produced straw hats in Tuscany in the past or still have connections to that world somehow.

Following, the project "Ti racconto in italiano" shall be illustrated. This is the result of the collaboration between the Istituto Centrale per i Beni Sonori e Audiovisivi (ICBSA, another office of the Ministry which concentrates its activity on the preservation and valorisation of audio and audiovisual heritage), l'Università di Siena (UNISI), l'Università di Siena per Stranieri(UNISTRA-SI) as well as the Biblioteca Nazionale Centrale of Florence (BNCF). This project is particularly interesting in this context as it bears agreat focus on indexing issues and, as it dates back to 2020, it adopts state of the art digital solutions.

These two projects have been chosen as interesting sample cases as they are both recent and quite unique in their sort. Cross-referencing between the worlds of library and archive databases is still relatively uncommon.

As concerns the "Archivi di paglia" project, it is necessary to illustrate at first the context in which the idea of such a research developed.

Archival records have a permanent significance for history, science and culture, as well as for the legal protection of individuals and legal entities. As such they can truly be considered to be a cultural asset.

Amongst the archival records that the Soprintendenza archivistica e bibliografica della Toscana safeguards there are, since the 1970's, business or company archives. Such archives bear witness to the history, capacities and vision of many big and small intrapreneurs in the local manufacturing industry. Tuscany has been famous since the 18th century for the production and processing of straw. The latter was used to realize the famous 'cappello di paglia di Firenze' (viz. the Florentine straw hat), giving work to hundreds of women employed in weaving straw into braids and hats. The tradition has continued somehow until today and some of the older firms are still active: in all, there are around 14 firms, many of which around a century old, which are still producing braids, hats and hat moulds. Their work and tradition has inspired the Museum of Straw in Signa[1], one of the main centres of the production of straw hats. It was in Signa that, for the first time, in 1714 Domenico Michelacci had the idea of starting a new kind of straw crop in order to obtain a thread that was particularly suited for weaving.

---

[1] https://www.museopaglia.it/ Accessed June 2021

JLIS.it

To celebrate the third centenary of the revolution in straw production and processing in the region after Michelacci's pioneering experiences, the Soprintendenza archivistica opted in 2014 to realize the census of the companies that were or are still active in the sector, to collect data on their archives and to organize and record a series of interviews with business owners, workers and furnishers active in this line (as well as their relatives) in order to have first-hand information on their way of life and work. The interviews (some on video as well) were intended for the Museum at Signa. One of the main aims of the project was also the publication online of the so-called SIUSA descriptions of the archival records of the firms involved. SIUSA is an acronym for the Sistema Informativo Unificato per le Soprintendenze Archivistiche (Unified Information System for the Supervision Agencies). It intends to be the primary access node to non State archival documents, both public and private, which are not kept by StateArchives.

The system describes the *archival fonds* according to a multi-level description; the *creators* (*bodies*, *people* and *families*) who produced the documents performing their activities; the persons or bodies who preserve (custody) the archives. General historical, administrative and archival information is provided as well, in order to allow a better comprehension of the context[2].

The aim of SIUSA is to assure the preservation and the knowledge of these sources and to provide access to them.

It was immediately clear from the perusal of the documents and the examination of the content of the interviews that it would have been essential to dispose of a controlled vocabulary focused on the world of straw processing. In such a way, it would have been possible to choose from a selection of terms in order to index content or to retrieve content through browsing or searching: thus, the SIUSA descriptions and any other works on the archives and interviews would have gained so much in sense, purpose and usability!

The potential of language as a meeting point between libraries, archives and museums was on the other hand a theme of reflection in those years for the Soprintendenza. So much so that in 2012 it had already created the basis of a collaboration with the Biblioteca Nazionale Centrale di Firenze through series of explorative letters and reciprocal declarations of intent.

For this reason, the Soprintendenza archivistica sought the collaboration of the Biblioteca Nazionale Centrale of Florence and specifically of the team behind the Nuovo soggettario[3]. The Nuovo soggettario viz. the New Subject Index, is the Italian subject indexing tool created by the National Central Library of Florence for the entire system of Italian libraries and, in particular, for those operating in the National Library Service.

Far from being a tool in use only in the Library world, the Nuovo soggettario was and still is open to contributions from other areas of knowledge management and is interoperable with databases of archives and museums as well as available in all standard formats and protocols.

Significantly the potential of this interaction between different worlds was explored on occasion of the conference organized in 2015 by ANAI, the Association of Italian Archivists[4], MAB *Musei Archivi Biblioteche. Professionisti del Patrimonio Culturale Toscana* as well as the Tuscan Region and hosted by the Soprintendenza itself. On that occasion Emilio Capannelli, one of the archivists

---

[2] https://siusa.archivi.beniculturali.it/cgi-bin/siusa/pagina.pl?RicLin=en Accessed June 2021
[3] https://thes.bncf.firenze.sbn.it/ Accessed June 2021
[4] http://www.anai.org/anai-cms/; http://www.mab-italia.org/ Accessed June 2021

JLIS.it

of the Soprintendenza, described the first experiences of collaboration between local archivists and librarians[5].

So, thanks to this collaboration with the National Library, Alessia Artini and Silvia Melloni, the two free lance archivists who worked on the straw archives project under the guidance of the Soprintendenza (and of the archivists Renato Delfiol and Luca Faldi in particular) produced a series of controlled vocabulary terms which have been accepted and adopted by the general Thesaurus that is the main component of Nuovo soggettario system[6]. These terms hence recur in the scientific production concerning the straw archives.

The idea was that of continuing the collaboration between archivists and librarians in order to create other carefully selected lists of words and phrases in order to tag units of information in other 'domains', such as the archival records of other manufacturing sectors. So far, though, this has not occurred, at least as concerns the Soprintendenza archivistica in Tuscany.

A more complex and recent project that develops significantly some of the themes touched by "Archivi di paglia" is that of "Ti racconto in italiano" (fig. 1). This one year long project which has been financed by the General Directorate of Libraries in the Ministry aims at promoting the collections of ICBSA as well as the knowledge of the Italian language and culture abroad. What is it all about? And, first of all, what is ICBSA?

Fig. 1.

---

[5] Capannelli 2016, 17-20
[6] Artini, Benelli and Melloni 2017, 9

JLIS.it

ICBSA – as stated before The Central Institute for Sound and Audiovisual Heritage[7] – was founded in 1928 and was once known as the State Discotheque. Its first collections consisted of a record collection entitled "The word of the Great", voices collected by Rodolfo De Angelis in the first half of the 1920s. Over the years, documents of folklore, music, history, theater, dance, cinema have been added to this initial nucleus, which represented the first Italian public sound heritage, recorded on the most different media, from the wax cylinders invented by Edison, to records, tapes, videos up to current digital media. ICBSA materials are public and available for consultation via OPAC (Online Public Access Catalogue) with the possibility of listening to the *incipit* of the digitized sound documents and consultation of the accompanying description.

At the beginning of 2020, a collaboration between ICBSA, and the University of Siena (UNISI) and the University for Foreigners of Siena (UNISTRASI) started in order to make a section of the "Historical Voices" collection available to the public.

The materials made available by ICBSA for the realization of the project are 35 interviews on audio files lasting an average of one hour each, carried out between 1983 and 2006. The interviews, chosen by Piero Cavallari (an ICBSA technician) involve prominent personalities from the world of Italian business, art and culture for a total of about 40 hours of recording.

In detail, the corpus of interviews comprises:

- 13 interviews with writers, intellectuals, actors, directors (1983-1989): Elio Filippo Accrocca, poet; Giorgio Bassani, writer; Attilio Bertolucci; poet; Giorgio Caproni, poet; Riccardo Cucciolla, actor; Margherita Guidacci, poet; Luciano Lucignani, film director; Luciano Luisi, poet, writer andreporter; Mario Luzi, poet; Ettore Paratore, Latin scholar; Guglielmo Petroni, writer; Luisa Spaziani, poet; Franca Valeri, actress. (fig. 2)

- 11 interviews with artists (1987-1988): Carlo Belli, intellectual interested in art, architecture, music, archeology, politics; Maria Lai, designer; Carlo Lorenzetti, sculptor in metal; Teodosio Magnoni, painter and sculptor; Elisa Montessori, painter; Alberto Sartoris, architect; Ruggero Savinio, painter and author; Toti Scialoja, painter and poet; Guido Strazza, experimental artist; Giuseppe Uncini, painter and sculptor in iron and cement; Renzo Vespignani, painter, illustrator, set designer and engraver.

- 11 interviews with entrepreneurs (2006): Pia Berlucchi, wine producer; Diana Bracco, health and diagnosys; Filippo Cerruti, tourism, maritime transport, events; Wanda Ferragamo, fashion; Vittorio Ghisolfi, plastic producer; Giorgetto Giugiaro, design; Sergio Giunti, book editor; Ernesto Illy, coffee producer; Steno Marcegaglia, steal producer; Loris Meliconi, house goods producer; Ottavio Missoni, sportsman and fashion designer.

---

[7] http://www.icbsa.it/ Accessed June 2021
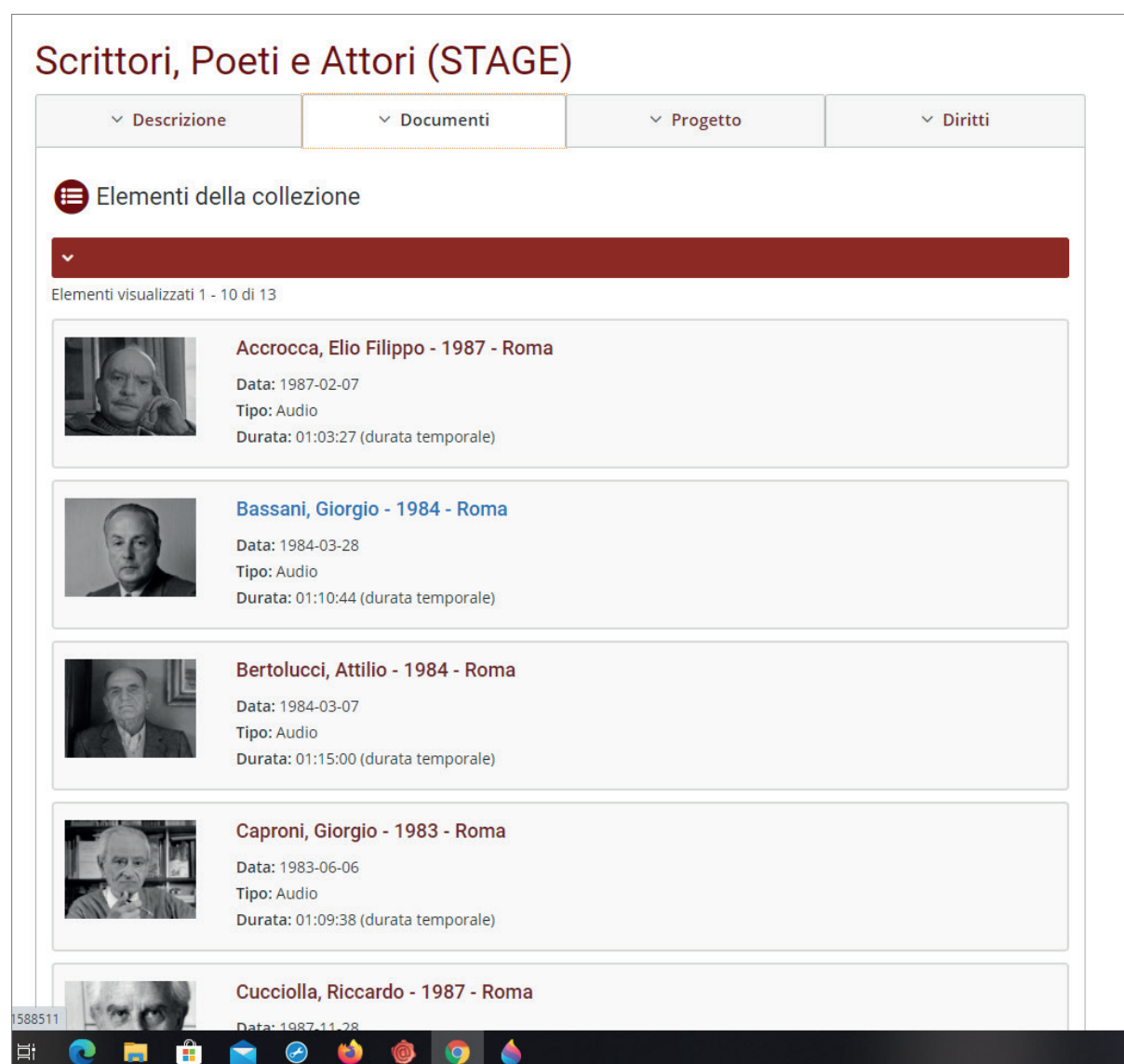
JLIS.it

Fig. 2.

As concerns UNISTRASI, ICBSA decided to activate a research grant in order to carry out the didactic adaptation of these sound materials to make them available online to teachers and students of Italian, Level B1-B2. The purposes were two: to bring foreign learners closer to intensive listening to encourage the learning of Italian, but also to open them a world of ideas, stories and emotions linked to some extremely interesting Italian personalities of the second half of the 20th century and through them to offer a cross-section of the Italian culture and society of this period. The work was accomplished by Elena Grifoni under the supervision of Pierangela Diadori (Full Professor of Didactics of Italian for Foreigners).

As Pierangela Diadori has noticed, it is difficult to think that a person with a mother tongue other than Italian and unfamiliar with the Italian culture of the second half of the twentieth century

JLIS.it

could listen to interviews lasting over an hour, or read the automatic transcription offered (honestly incomprehensible in many points): not even an Italian native speaker would do it.

The goal was therefore to create a battery of metadata, which could be freely surfed online and associated with the audio files and written files relating to each interview. It all had to be in the form of comprehension or completion exercises with closed answers, to be carried out in relation to the 'listening or reading'of the texts provided, together with keys for self-learning (figs. 3-4).

Fig. 3.

JLIS.it

Fig. 4.

The format had to be such as to allow the insertion of these sets of exercises on the platform, in order to make access faster and offer immediate feedback with keys to the answers. Also a glossary has been realized.

If one thinks of the shortage of didactic audio materials, having this instrument available anywhere in the world online, with the advantage of being free of charge and the possibility of accessing authentic cultural contents, it is easy to understand the importance of this operation[8].

UNISI's contribution to the project consisted, instead, in the indexing of the text in order to make it interesting for a wider number of potential users and perusable for research purposes. In this

---

[8] Diadori 2021

JLIS.it

case as well a research grant was activated: the work was accomplished by Cecilia Valentini under the supervision of Silvia Calamai (Associate Professor of Glottology and General Linguistics).

As Valentini and Calamai have pointed out on occasion of the presentation of the project at the CLARIN Annual Conference 2020[9], the main objective of "Ti racconto in italiano" was to create tools facilitating users search through the collection. Therefore, finding aids such as indexes and thesauri have been realized.

Indexing has been done classifying each segment of the documents at regular time intervals with a label (fig. 3).

The terms used as labels consist of key words and controlled vocabulary and are structured in a specifically created thesaurus, viz. a specialized vocabulary of hierarchically listed words and phrases that indicates a preferred term among synonyms and shows relationships between terms. The use of a thesaurus facilitates retrieval of information and ensures greater consistency in the indexing of documents.

The thesaurus used is naturally that of the Nuovo soggettario[10]. The National Central Library of Florence is in fact a partner of the project. Therefore, Anna Lucarelli and her team have followed all phases concerning the choice of the terms and have authorized the use of new terms proposed by Cecilia Valentini and which have now entered the Nuovo soggettario.

Indexing has been done, mainly by Stella Montanari via AVIndexer, a software developed by Davide Merlitti (Informatica Umanistica, Pisa)[11] that makes use of SKOS, the Simple Knowledge Organization System recommended by W3C. Also the DublinCore Metadata set has been exploited (figs. 5-6).

Once indexed, the records as well as the thesauri will be published on the internet portal "Ti racconto in italiano" managed by ICBSA. The process is in progress at the moment. This platform is shaped on the digital library *Ti racconto la storia* online since September 2018[12]. The latter was conceived by the General Directorate for Archives and the Central Institute for Archives in order to promote the knowledge and use of collections of oral testimonies, stories of life and other audio and audiovisual documentation produced on both analogue and digital media and stored in public institutions, research centers and private associations.

Once online, the perusal of the platform[13] will clearly show how the resources of the Semantic Web have been exploited in order to create a framework for creating, managing, publishing and searching semantically rich metadata for web resources.

---

[9] https://www.clarin.eu/content/programme-clarin-annual-conference-2020 Accessed June 2021
[10] https://thes.bncf.firenze.sbn.it/ Accessed June 2021
[11] http://www.informaticaumanistica.com/open-source/avindexer Accessed June 2021
[12] https://www.tiraccontolastoria.san.beniculturali.it/ Accessed June 2021
[13] The foreseen address is: https://tiracconto.icbsa.it/

JLIS.it

Fig. 5.

JLIS.it

Metadati MPEG-7

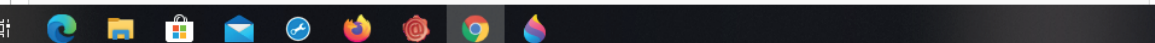|	
Titolo	Bassani, Giorgio - 1984 - Roma
Riassunto	Dopo la nota biografica e bibliografica a cura di Eugenia Tantucci, Bassani passa lettura di versi tratti da varie raccolte, di cui descrive brevemente le forme e i riferimenti. Durante la lettura ribadisce la centralità dell'antifascismo, dell'ebraismo, delle memorie d'infanzia nella sua poetica. Insiste sulla sostanziale uguaglianza fra le sue opere in versi e in prosa e conclude leggendo il capitolo 9 de Il giardino dei Finzi Contini.
Produttore	Discoteca di Stato
Intervistatore	Tantucci Eugenia
Intervistato	Bassani Giorgio
Luogo	Roma
Data	1984-03-28
Genere	Intervista
Forma	Serie
Soggetto	letteratura / Italia / 1980-1989 / poesia / interviste
Lingua	it-IT
Durata	01:10:44
Parole chiave (SKOS)	Bologna / Università degli Studi di Bologna / Longhi, Roberto / Tantucci, Eugenia / Famiglie / Ferrara / Città / Ispirazione poetica / Attaccamento / Letteratura / Politica / Attività clandestina / Giovani / Intellettuali / Partito d'azione / Arresto / Antifascismo / Armistizio dell'8 settembre <1943> / Roma / Pubblicazione / Poesia / Bassani, Giorgio, Storie dei poveri amanti e altri versi / Bassani, Giorgio, Te lucis ante / Bassani, Giorgio, Un'altra libertà / Bassani, Giorgio, L'alba ai vetri / Redattori / Periodici / Botteghe oscure <Periodico> / Paragone <Periodico> / Bassani, Giorgio, Cinque storie ferraresi / Bassani, Giorgio, Gli occhiali d'oro / Bassani, Giorgio, Il giardino dei Finzi-Contini / Premio Viareggio / Bassani, Giorgio, Dietro la porta / Bassani, Giorgio, L'airone / Premio Campiello / Bassani, Giorgio, Le parole preparate e altri scritti di letteratura / Bassani, Giorgio, L'odore del fieno / Racconti / Saggi / Bassani, Giorgio, Epitaffio / Bassani, Giorgio, In gran segreto / Bassani, Giorgio, Il romanzo di Ferrara / Dediche / Bassani, Giorgio, In rima e senza / Premio Bagutta / Dortmund / Premio Nelly Sachs / Traduzioni / Narrativa / Sintassi / Lessico / Risentimento / Dolore / Vita / Storia / Solitudine / Discriminazione razziale / Religiosità / Laicismo / Emarginazione / Società / Contestazione / Dubbio / Antenati / Libertà / Bassani, Giorgio / Rima / Poeti / Versi / Alessandrini / Dodecasillabi / Settenari / Strofe / Iscrizioni / Epitaffi / Lapidi / Cimiteri / Case / Jahier, Piero /

Fig. 6.

JLIS.it

References

Artini, Benelli and Melloni 2017. *Archivi di paglia. Gli archivi del distretto industriale della paglia in Toscana*, edited by Alessia Artini, Angelita Benelli, Silvia Melloni, Firenze: Edizioni Polistampa.

Capannelli, Emilio 2016. "Prime esperienze di collaborazione tra archivisti e bibliotecari" In *Il nome delle cose. Il linguaggio controllato come punto di incontro tra archivi, biblioteche e musei. L'esperienza del Gruppo linguaggi di MAB Toscana*, edited by Francesca Capetta, 17-20. Accessed June 2021. http://www.ilmondodegliarchivi.org/images/Quaderni/MdA_Quaderni_n1.pdf

Diadori, Pierangela 2021. "Ti racconto in italiano: voci del '900 per imparare l'italiano L2. Progetto ICBSA-UNISTRASI di didattizzazione di interviste sonore a personalità del '900." In La Nuova DITALS risponde 3, edited by Pierangela Diadori, Caterina Gennai, Elena Monami, in press. Roma: Edilingua.

JLIS.it

The bibliographic control of music in the digital ecosystem. The case of the Bayerische Staatsbibliothek (BSB)

Klaus Kempf[(a)]

a) Independent consultant – formerly Bayerische Staatsbibliothek

ABSTRACT

The BSB's music department (entrusted since 1949 with the management of the national information service on music) is one of the largest music libraries in the world in terms of the size and quality of its collection, but also in terms of the breadth and depth of its collection acquisition policy. The various materials are widely catalogued and indexed in a very articulate way, using a wide range of catalogues and according to specific rules. The BSB currently uses the RDA and MARC21, according to national policies.

The Gemeinsame Normdatei (GND), the authority files of the German-speaking library world, are used both in cataloguing and in subject classification. The GND is nowadays used even outside the library world by archives, museums and other kinds of institutions, as well as for the cataloguing of websites.

The BSB participates in the RISM (Répertoire International des Sources Musicales) international online catalogue of music sources, and, together with the Staatsbibliothek zu Berlin, manages its OPAC.

The presentation will describe these projects, as well as the cataloguing workflow, the application of the RDA in specific cases, the special rules (and cataloguing system) for personal archives and musical legacies (RNA), and finally the futuristic service 'musiconn'. This last service is included in the national service for music information Fachinformationsdienst Musikwissenschaft and has been developed by the BSB: it offers the possibility to search by melody, as part of a project based on Optical Music Recognition (OMR), a software tool that allows automatic recognition of compositions after they are printed.

KEYWORDS

Cataloging music sources in Germany; German authority files in musicology and music sources; Digitization of music sources.

JLIS.it

The Bayerische Staatsbibliothek as a big music research library

Subsection title

The Bayerische Staatsbibliothek (BSB) is not only the central state library (national library) of the Free State of Bavaria, but also or in particular a big research library. It disposes of world wide well known and recogniced special collections in a couple of science disciplines. One of them regards music and musicology. The library hosts in its stacks 455.000 music editions, 72.000 music manuscripts, 330 composer archives (personal papers), 93.000 non book material/sound carriers, in particular discs and CDs, 164 000 books and journals about music and musicology. Since 1949 the BSB is part of the national special collection progamme, especially music, cofinanced by the German Research Society (DFG) and since 2014 the library is responsable for music and musicology within the framework of the Specialised information services programme (FID) also cofunded by the DFG.

Cataloging & metadata management in music and musicology collections

Following the principle that cataloging aims traditionally on two major objectives: on the one hand side on the specifica of the object/material regarded: on the other hand side on the needs (and desires) of the (potential) user leads in the field of music and musicology – at least in the German speaking world – to an especially varied and contemporarily particularly profiled cataloging/metadating in a relatively wide range of different catalogs.

Since the introduction of the online catalogs (OPAC) the German libraries use more and more the in the meanwhile well established standards,

- formed by the national/international cataloging rules together with the special guidelines for music (in former times RAK /Musik now a days RDA and the music specific guidelines – https://wiki.dnb.de/display/RDAINFO/Arbeitshilfen); and
- standardised data formats (MAB2 and MARC 21; in the case of personal papers also EAD);
- today libraries use for cataloging, and subject heading (the same) authority files. In Germany in both cases they use the Integrated Authority File (GND).

The integrated authority file (GND)

The Integrated Authority File (Gemeinsame Normdatendatei – GND: https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html;jsessionid=935B36EDCD89249E62A%201BA3000574759.internet531) is a service facilitating the collaborative use and administration of authority data. These authority data represent and describe entities, i.e., persons, corporate bodies, conferences and events, geographic entities, topics and works relating to cultural and academic collections. Libraries in particular use the GND to catalog publications. However, archives, museums, cultural and academic institutions, and researchers involved in research projects are also increasingly working with the GND. Authority data make cataloging easier, offer definitive search entries and forge links between different information resources. Every entity in the GND features a unique and stable identifier (GND ID). This makes it possible to link the authority data

JLIS.it

with both each other and external data sets and web resources. This results in a cross-organizational, machine-readable data network.

Cataloging of music editions, books, and audio/sound carrier in Germany

Cataloging of music editions, books and audio/sound carriers in Germany is normally done on a regional level via the academy library system dominating regional and intraregional library networks, however periodicals are cataloged in a nationwide, even transnational database. In the case of the BSB music editions, books (monographs) and audio/sound carriers are cataloged in the Union Catalog (Verbundkatalog – B3kat) of the Bavarian Library Network (BVB): <https://www.bib-bvb.de/>. But the periodicals, journals, year books and so on, are cataloged – like the other libraries are doing – in the German National Periodical Catalog (ZDB): <https://www.zeitschriftendatenbank.de/startseite> on which are participating also the library systems of Austria and the German Switzerland. By adding a shelfmark at the cataloging record in the union catalog and the national periodical catalog the concerned record is replicated/duplicated – in real time – in the regarded local OPAC: <https://opacplus.bsb-muenchen.de/metaopac/start.do>.

In a different way are handled the music sources. They are primarily cataloged worldwide in cooperation via RISM <https://rism.info/index.html>. The Répertoire International des Sources Musicales (RISM), International Inventory of Musical Sources, is an international, non-profit organization that aims to comprehensively document extant musical sources worldwide: manuscripts, old music editions, writings on music theory, and libretti that are found in libraries, archives, churches, schools, and private collections. The RISM Catalog of Musical Sources contains over 1.2 million records and can be searched at no cost. RISM was founded in Paris in 1952 and is the largest and only global organization that documents written musical sources. RISM records what exists and where it can be found.

The cataloging happens decentrally via an international cooperation following specific cataloging rules which are primarily oriented on musicological criteria. This catalog in the case of handwritten material offers access to its content also via the cataloging of the so called music incipits. Moreover, this catalog is using an own rather detailled (meta)data format. At least the German editorial staff is also using systematically the (German) Authority File, the GND. The central cataloging tool, the data base, is hosted by the Staatsbibliothek zu Berlin (SBB). The (RISM) OPAC is managed and maintained by the BSB. The BSB-OPAC is updated every half a year with the new records cataloged in the RISM-Catalog via the Bavarian Network Union Catalog (B3Kat).

Cataloging of personal papers (composer archives), publisher archives and manuscripts

Kalliope is a Union Catalog for collections of personal papers, manuscripts, and publishers' archives and the National Information System for these material types (https://kalliope-verbund.info/de/index.html). Founded by the Berlin State Library – Prussian Cultural Heritage with financial support from the German Research Foundation (DFG) in 2001, Kalliope superseded the Central Register of Autographs (Zentralkartei der Autographen, ZKA), which was established in 1966.

JLIS.it

The joint cataloging in Kalliope is based on established archival and librarian description and cataloging guidelines and relies heavily on authority control processes. Kalliope is therefore not just another data aggregation service, but rather a digital environment that establishes and provides new instruments and processes to create, modify, and to access data about personal papers dispersed in many libraries, archives, and museums. But using Kalliope means parallel cataloging and parallel offer of data access. Until today there is no connection, neither an interface nor a data transfer (replication) between the Kalliope data base and the union catalogs of the various regional library networks or the single local OPACs.

Kalliope: History, Development, State of the Art

The initial data base of Kalliope was formed by 1,2 million catalog cards of the ZKA that had been provided by 450 institutions over a period of more than 30 years. The conversion of these cards into a machine-readable format was completed in 2006. Moreover, the catalog service was extended step by step to provide access to collections of personal papers in Austria and Switzerland as well as personal papers of persons from German speaking countries kept in libraries and archives abroad, particularly in the United States of America. Since 2001 cultural heritage organizations can make full use of a client-server based cataloging application including full access to the Integrated Authority File (Gemeinsame Normdatei, GND) of the German National Library. The Union Catalog takes an active role in the operation of this national cataloging resource and adds greatly to it by identifying entities that are only known via unique materials as are described in Kalliope.

The cataloging client conforms to the German Guidelines for the description of personal paper and manuscript collections (RNA – Regeln für die Erschließung von Nachlässen und Autographen) which are in turn compatible with ISAD(G) – General International Standard Archival Description. As of May 2015, 102 organizations use the Kalliope cataloging client – compared to 54 in 2010. Additionally, standardized data (EAD – Encoded Archival Description) from local applications can be made available for retrieval in Kalliope. Currently the database provides access to 19,300 collections with a total of more than 3 million units of description originating from more than 950 institutions, including letters, manuscripts, personal documents, albums, diaries, lecture notes, photographs, posters, movies, screenplays, music editions and even some famous ringlets.

The database includes around 600,000 name records, 253,000 of which describe individualized persons distinguished by a unique identifier of the GND, and more than 90,000 records of corporate bodies, with 24,000 of these having a unique identifier of the GND.

Kalliope and bibliographic control

The Kalliope Union Catalog is committed to comply with standards (guidelines, file formats, authority files, ISO norms) of the library and archival community: Encoded Archival Description (EAD): XML schema for encoding archival finding aids; Encoded Archival Context – Corporate Bodies, Persons, and Families (EAC): XML schema for encoding (archival) authority records; GND – Integrated Authority File of the German National Library: uniquely referenced vocab-

JLIS.it

ulary for entities such as persons, corporate bodies, places, and subject headings Guidelines for the Description of Personal Paper and Manuscript Collections (Regeln für die Erschließung von Nachlässen und Autographen, RNA); ISO 15511: International Standard Library Identifier and Related Organisations (ISIL): unique identifier code for scientific and cultural heritage organizations; ISO 3166: Codes for Names of Countries, dependent territories, special areas of geographical interest and their principal subdivisions: used for asigning persons and corporate bodies of the integrated authority file to main geographical area; ISO 629-2: Codes for the representation of names of languages – Part 2: Alpha-3 code: used to describe the language within a unit of description. The Guidelines for the Description of Personal Paper and Manuscript Collections are well established and applied by libraries, archives, museums, and similar organizations in Austria, Germany, and Switzerland, and are compatible with the principles of archival description outlined in the General International Standard Archival.

The new Special information service (FID) for music/musicology – musiconn

The introduction of information portals or platforms is a web conform way to put order and structure in rather heterogenous data, but concerning the same scientific discipline; in particular they offer normally a discipline specific unique search possibility on heterogenous information sources. In German this way of presenting information is often called Sekundär-erschließung. The service musiconn.search (https://www.musiconn.de/) offers access to 19 relevant data bases/catalogues and other, also fulltext online sources with 6,5 millions single items. A special service of musiconn is the so called melodies search, the musiconn.scoresearch. The prototypical application was developed by the BSB itself, promoted and financially supported by the DFG. The software tool is based on the principles of Optical Music Recognition (OMR) and allows the automatical recognition of melodies in selected digitized music sheets. Actually the melody search is possible in the compositions of the following composers (https://scoresearch.musiconn.de/ScoreSearch/about).

Conclusion

The start of the online cataloging pushed the standardization in general and involved a consequent usage of authority file controlled terms in cataloging as well as in subject heading. This applies in principle also to information material regarding music and musicology. In the field of music and musicology cataloging traditionally there is a strong input from the research community itself. The organizational platform for that is RISM, an international body which has established just in the early 50ies its own cataloging database. Today the database is a publicly accessible catalog for the registering (and cataloging) of music sources, in particular manuscripts and old music editions as well as libretti. The RISM-Catalog has its own rules and is based on its own data format. This catalog offers access to its content also via the cataloging of the so called music incipits.

In the German speaking world existing online platform for the cataloging of personal papers & autographs and giving access to them in the internet, called Kalliope, is also used for the catalog-

JLIS.it

ing of relevant material in music and musicology. Last but not least the recently introduced special information service (FID) for music, musiconn, offers with musiconn.scoresearch the possibility to find melodies in selected and via the portal accessible digitized music editions. Score search is still a work in progress and not error-free, but it is a decisive step towards a machine learning approach in the field of musicology. It is similar to a full text search in text based sciences and in a near future it can become an very interesting service also for the average user.

JLIS.it

Riviste digitali e digitalizzate italiane (RIDI): a reconnaissance for the national newspaper library

Fabio D'Orsogna[(a)], Giulio Palanga[(b)]

a) Biblioteca nazionale centrale di Roma, http://orcid.org/0000-0001-9578-8715
b) Biblioteca nazionale centrale di Roma, http://orcid.org/0000-0001-9737-2529

ABSTRACT

The article presents a reflection born from a reconnaissance (named RIDI, Riviste digitali e digitalizzate italiane) launched in December 2019, on online open access journals and digitalization of previously printed publications, which are not always considered as unitary bibliographic elements. It highlights the increasingly urgent need to offer its users not only the physical heritage of the library but also the entire world of open-access digital publications available on the web. Starting from an overview of the state of the art of Italian open access periodicals, both digital natives and continuations or parallel editions of previously printed publications, it offers some examples of bibliographic records already present in the national OPAC of SBN (Italian union catalog), related to publications available with both printed and digital editions. It illustrates the main Italian and international digital libraries, highlighting the problems of coordinating the various initiatives to improve the quantitative and qualitative offer of products. The directory, from which a database integrated with the portal of the Digital Newspaper Library of the National Central Library of Rome will originate, will allow direct access to resources through multiple search fields. The prototype of a super-record of the Work will provide all the elements for the standardization of images, data, metadata, bibliographical histories of publications, to build the national digital newspaper library of the future.

KEYWORDS

Digital libraries; Digital newspaper library; Open access serials; RIDI; National Central Library of Rome.

JLIS.it

This paper takes its cue from a reconnaissance called RIDI (Riviste digitali e digitalizzate italiane) launched in December 2019, on those bibliographic realities often not considered in a unified way on our OPACs, such as open access online journals and digitalization of previously printed publications. The need, more than 20 years after the first digitization projects that involved Italian libraries, is to offer its users not only the physical heritage of its library but also the whole vast world of open access digital publications available on the web.

The paper will illustrate the state of the art on Italian open access periodicals, either digital natives, continuations, or parallel editions of previously printed publications, then it will propose some examples of bibliographic records already present in the Italian national catalog, the SBN (Servizio bibliotecario nazionale) OPAC,[1] related to publications with both printed and digital editions, then it will provide a prototype that can provide the elements for a qualitative standardization of images, data, metadata, bibliographic histories of publications, to build the national digital newspaper library of the future.

Better late than never

In the Wiki on Open Access in Italy, a portal that records news and information about the movement at the national and international level, under the heading *Riviste italiane OA* (Italian OA Journals), this communication appears: "At this time the requested page is empty. You can search for this title in other pages of the site or search in related registries, but you do not have permissions to create this page".[2]

This first Italian directory, which does not presume to be exhaustive, arrives with some delay and tries to fill a gap. It is now about 20 years that the main faculties and university departments in the world have begun to organize themselves to offer their journals online. In 2000, the Cato Institute, a temple institution of US liberalism, dedicated its annual conference to the question of which of the two paradigms – intellectual property or open access – would dominate the economy of the future.[3] In 2003, many scientific institutions signed the Berlin Declaration on Open Access to Scientific Literature. In Italy, in November 2004, the Berlin Declaration was followed by the Messina Declaration, joined by about thirty universities.[4]

Universities began to organize themselves by creating dozens of university presses. In 2009 the UPI Coordination was established, which in 2018 became the *Associazione Coordinamento delle University Press Italiane.*[5]

---

[1] https://opac.sbn.it/opacsbn/opac/iccu/free.jsp

[2] https://wikimedia.sp.unipi.it/index.php?title=Riviste_italiane_OA

[3] Carlo Formenti, *Corriere della sera*, 20 novembre 2000, p. 27; see also http://www.treccani.it/vocabolario/open-access_%-28Neologismi%29/

[4] Bologna, Brescia, Calabria, Firenze, Foggia, Genova, Insubria, Lecce, Messina, Milano, Milano Bicocca, Milano Politecnico, Milano Vita-Salute San Raffaele, Modena, Molise, Napoli Federico II, Napoli L'Orientale, Napoli Partenope, Padova, Palermo, Parma, Piemonte Orientale, Roma LUMSA, Roma Tor Vergata, Roma III, Siena, Torino, Trieste, Trieste SISSA, Tuscia, Venezia IUAV, and Istituto Italiano di Medicina Sociale di Roma.

[5] The association aims to study and deepen the issues related to the positioning, the function, and the promotion of university publishing and popular science as well as the possibility to participate in national and international calls for funding of publishing projects. 13 university publishing houses publish 25 open access journals.

JLIS.it

Reasons for growth

We are still far from the total replacement of the printed page by the web page, but the steady growth in the number of magazines appearing online is no less real.[6] The reasons for this success are very simple: plenty of space reduced publication costs and, above all, ease of access anywhere with just the availability of a network connection.

Underlying the success of the Open Access Initiative are two instances:

1. increase dissemination, visibility, and impact of scholarly literature through publication in open, online, institutional, and disciplinary repositories;
2. to counteract the rising prices of academic journals with alternative models of scholarly communication.

For many small businesses, bearing the economic burden of printing and shipping magazines has become unsustainable and is often the motivation to publish only in digital format. This transformation, feared by many, which represents a surrender to the affordability of digital, often also allows a qualitative leap and a broadening of the horizons of publications. Online publication can enhance the characteristics of periodicals and allow readers to navigate the texts in a simpler, more agile, and sometimes interactive way.

There are two models for sustaining management costs and remaining adherent to the philosophy of free access; the model centered on financing by consumers of content (demand-side) and that financed by content producers through sponsorship, donations, fundraising (supply-side). The main supply-side model is that of the Article Processing Charge (APC), better known as the author-pays model, which provides for the payment by the authors of articles accepted for publication of a contribution, which can reach in some cases up to $ 2,500, while for the contributors of articles from poor or developing countries, the publication is free.[7]

Legal deposit of digital resources in Italy

Legal deposit, i.e. the compulsory delivery of publications to depository institutions by the subjects envisaged by Italian Law no. 106 of April 15, 2004, and Presidential Decree no. 252 of May 3, 2006, is the regulatory instrument that allows the collection and preservation of the various publications in national and regional archives. The law also deals with native digital publications (born-digital).

Two significant experiences were born as a result of the law.

CNR SOLAR (Scientific Open-access Literature Archive and Repository) is a database of scientific publications, established in 2006, aimed at creating an archive of Italian products of science and research, using also the Legal deposit of digital publications. In the context of the mission entrusted to the CNR Central Library by the Law 2004/106 and by the Presidential Decree 2006/252, the

---

[6] The Directory of Open Access Journals (DOAJ) listed 2,100 academic-level journals in 2006; as of April 3, 2021, there are 16,146.

[7] A clear and comprehensive account of the costs of the Open Access publication process can be found in Technical report #1 (2018) from CNR Bologna: Mangiaracina, Silvana and Cristina Morroni. 2018. *Quanto costa l'accesso alle pubblicazioni scientifiche nell'era dell'Open Access? : una prima analisi delle pubblicazioni nel CNR*. Bologna: Biblioteca Area della ricerca di Bologna CNR. https://zenodo.org/record/1247497#.XoC-JKPOPkU

JLIS.it

legal deposit is aimed at constituting the Italian archive of scientific publications and at realizing national bibliographic services of information and access to the documents subject to legal deposit. Legal deposit in SOLAR is realized through:

1. self-archiving by the author(s), who must make sure of the actual conditions of use and dissemination of the version of the deposited work, previously agreed upon with the publisher and/or producing institution;

2. specific agreements between the CNR Central Library and the publisher and/or the producing institution of the publications. In this case, the deposit may be made by the Central Library itself or by the publisher/producing institution.

The resources in SOLAR can be full-text open access or limited access, i.e. the metadata are still accessible, while it is necessary to contact the CNR Central Library for full-text resources.

Magazzini Digitali is the Italian project for digital legal deposit, launched on July 14, 2011, with the signing of an agreement between the Ministry of Cultural Heritage and the Presidents of the most representative associations of the publishing industry: AIE, FIEG, USPI (later joined by MEDIACOOP and ANES).

The purpose of the agreement was to promote the experimentation of the legal deposit of born-digital works in the National Central Libraries of Rome and Florence and, limited to the backup copy, in the Biblioteca Nazionale Marciana of Venice.

The experimentation lasted 3 years, starting in 2012. After this period, a shared and efficient system of legal deposit should have been outlined and, in particular, the procedures related to digital works should have been defined through the issuing of a specific regulation.

The trial ended on December 31, 2014. Magazzini Digitali continued to receive subsequently few publications covered by the agreement in the 2012-2014 Conventions, receiving at sperimentazione@depositolegale.it requests for voluntary membership, while waiting for a final regulation and trying to cope with requests based on the few resources available.

The budget is insufficient, as is, more generally, the response of Italian cultural institutions to the preservation of this type of publications, which will inevitably be lost if no concrete and adequate action are taken to deposit them in national archives as is the case for printed publications.

Contents and purpose of RIDI

RIDI (Riviste digitali e digitalizzate italiane) is a repertory conceived as a work in progress that already contains the bibliographic records of about 12,000 Italian journals, compiled according to the standards of the SBN cataloguing guide,[8] with their URIs,[9] available on the Internet for free access. All journals that require subscription and registration for a fee are excluded.

There are two main reasons for this choice: the first is practical and is based on the consideration that online journals now represent an enormous quantity, probably more than that of printed journals, which makes bibliographic control almost impossible, as Mauro Guerrini predicted in

---

[8] https://norme.iccu.sbn.it/index.php/Guida_moderno
[9] https://it.wikipedia.org/wiki/Uniform_Resource_Identifier

JLIS.it

1999.[10] The second is more exquisitely librarian: both from the cataloguing point of view since it gives a standardized account of bibliographical descriptions that would otherwise be absent from the web and, above all, from national and local OPACs of resources that are unknown to catalogs; and from the point of view of the preservation of printed copies of dual-track publications (paper and online), since it would be possible to exclude from ordinary consultation all those resources that are freely available on the Internet and of which information has been given in catalogs.

The repertory is currently complete only for digital journals and is being completed for the part concerning journals digitized from paper format. The final goal will be to create a single repertory containing also the ever-growing world of resources born in print and digitized later, as a result of public and private digitization campaigns in recent decades. The digital recovery of a printed past, among other things, is present in many journals that are entering open access after a long paper season and represent an attempt to progressively provide all the published material in a single digital archive. Think, for example, of what Banca d'Italia has done in the last 10 years (it has 97 open access publications in its portal) which has made an enormous recovery of its historical publications.[11]

The search for digitized journals began in April 2020. The work will involve the analytic cataloguing of 73 Italian digital libraries surveyed. As for the type of resources, RIDI includes:

a) Italian native digital journals, which are one-tenth of the total;
b) journals published in mixed form, in print and online. Of the publications of this second type, the description of the printed part has also been given, highlighting all the connections between the two forms of publication;
c) digitized Italian journals.

Intending to find titles even outside the academic circuits, we have therefore given an account of the relations, more and more numerous and frequent, between printed publication and open access publication within the history of the same publication. This allowed us to adequately reconstruct the historical evolution of many journals, also from a cataloging point of view, to offer the OPAC, in the case of Italy the SBN OPAC, the possibility of providing adequate information on their publishing history and to start the cataloging of the online issues in SBN, both by intervening in the area of notes and URIs[12] and by creating new bibliographic descriptions linked to the descriptions of the printed editions. During the editing of this catalog, numerous bibliographical descriptions of online resources not yet present were created on the SBN OPAC.

Take the case of *Giornale di gerontologia*, a prestigious journal published for sixty years by the Italian Society of Gerontology and Geriatrics. In 2013, it ceased its print publication. A laconic note informs SBN OPAC users that since 2014 it is published only online. Actually, the journal

---

[10] "La proliferazione incontenibile delle basi di dati ad accesso remoto rende evidente come sia oggi più che mai illusorio il controllo bibliografico universale [...] la biblioteca può pensare di descrivere solo le risorse elettroniche di proprio interesse [...] selezionando le risorse in modo piuttosto stretto". Guerrini, Mauro, "Catalogare le risorse elettroniche: lo standard ISBD(ER)", *Biblioteche oggi*, 17 (1999), no. 1, 62.

[11] See https://www.bancaditalia.it/pubblicazioni/relazione-annuale/index.html for the annual reports of Banca d'Italia governors from 1894 to 2019.

[12] In this regard, see the ICCU note that established the rules for including in the General Content Notes of the ISBD the link to the digitized copy of the copy not owned by the operating library. http://polonap.bnnonline.it/index.php?it/21/news-ed-eventi/46/link-alla-copia-digitalizzata-dellesemplare-non-posseduto-dalla-biblioteca-operante-regole-per-linserimento

JLIS.it

only retrieves online a few previous years and since 2016 it changes its title. Users have no news of this. RIDI offers this fundamental information to reconstruct the entire bibliographic history of the periodical.

---

*Giornale di gerontologia : organo ufficiale della Società italiana di gerontologia e geriatria. – Anno 1, n. 1/2 (gen.-feb. 1953)-anno 61, n. 6 (dicembre 2013). – Firenze : L. Macrì, 1953-2013. – 61 volumi : ill. ; 25 cm. ((Mensile; poi bimestrale. – Il formato varia in 30 cm. – La casa editrice varia: Pisa : Pacini. – BNI 1953-5821. – ISSN 0017-0305; poi 0367-4533. – Dal 2014 solo on line. – CFI0353910

Ha come supplemento: *Giornale dell'arteriosclerosi

*Giornale di gerontologia : organo ufficiale della Società italiana di gerontologia e geriatria. – Anno 58, n. 1/2 (gen.-feb. 2010)-anno 63, n. 4 (dicembre 2015). – Pisa : Pacini, 2014-2015. – 34 File PDF. ((Bimestrale; trimestrale nel 2015. – ISSN 2035-021X. – Disponibile in Internet all'indirizzo: http://www.jgerontology-geriatrics.com/issue/archive

Continua con: *JGG : *Journal of gerontology and geriatrics

*JGG : *Journal of gerontology and geriatrics : official journal of the Italian Society of gerontology and geriatrics. – Vol. 64, 01 (2016)-. – Pisa : Pacini, 2016-. – File PDF. ((Trimestrale. – ISSN 2499-6564. – Disponibile in Internet all'indirizzo: http://www.jgerontology-geriatrics.com/issue/archive
Autore: Società italiana di gerontologia e geriatria

Soggetto: Geriatria – Periodici; Gerontologia – Periodici
Classe: D618.97005

Giornale di gerontologia *bibliographic record on RIDI*

RIDI is ordered alphabetically by title. The source of the bibliographical information is CAPUS (Catalogo delle Pubblicazioni in Serie possedute dalla Biblioteca nazionale centrale di Roma). [13] In order to give full visibility to this repertory and to expand its search possibilities, it will be necessary to create a database, which will allow direct access to the resources through multiple

---

[13] CAPUS is a catalog edited by Giulio Palanga and published in 2019, containing the complete collection of all periodicals and newspapers owned by the National Central Library of Rome. Divided into twelve volumes, the first two volumes contain the index of titles and the index of authors and subjects and constitute the guide by which to navigate the catalog knowing a title, an author, or a subject to search for. The other ten volumes contain the bibliographical records of over 72,000 periodicals, divided into 53 sections, which contain a unique alphanumeric code that refers to a single access point in the catalog, with all the history and editorial changes of the publication, without the need to navigate through the various titles that periodical publications often adopt. http://www.bncrm.beniculturali.it/it/325/archivio-news/3259/

JLIS.it

search fields.[14] All the descriptions, however, already allow a hypertextual link to digital or digitized resources.

This first compilation of the catalog, completed on April 7, 2020, includes 11,706 bibliographic records. We started by retrieving information from CAPUS, where over twelve years of editing, links, and URIs with online publications of printed journals were gradually reported. This first reconnaissance has allowed us to find the journals contained in the 66 main Italian open access publishing platforms that have been analytically catalogued, verifying the correctness of the URI links, leaving out those no longer traceable on the web. The vast majority of these titles are also present in the two most important international sources, the ISSN portal with 1,028 titles,[15] the DOAJ (Directory of Open Access Journal) with 461 titles, [16] and, for Italy, Magazzini digitali with 113 titles. [17]

Open access, digitization, and bibliographic control

Online publications are often accompanied by the digital retrieval of issues published in print. Sometimes, this can happen by chance, but it is now possible to reconstruct and document the history of a publication through the various phases of its editorial policy, which almost always start from the printed text and end with the online publication and the digitalization of previous years.[18]

See, for example, the digitization of the entire archive of Radiocorriere, a weekly magazine that was the official organ of RAI for seventy years, from 1925 to 1995. With all the schedules and articles of the newspaper, it is possible to reconstruct the bibliographic (and political) history of the publication. An enormous amount of unpublished material, which represents a unique testimony and an exclusive source for contemporary historiography, not only of the media. It is one of the treasures recovered by Teche RAI and made available to the network free of charge.

---

[14] For example, in addition to the title, one could include search fields for author, subject, DDC, and provide a permalink of the resource and a permalink of the description in the catalog, as well as the holdings of the printed resource and the holdings of the online or digitized resource.

[15] For the alphabetical list of Italian periodicals see: https://portal.issn.org/?q=api/search&search[]=MUST=country=I-TA&search_id=7564722&sort=sort.title. Not all titles belong to actual periodicals. As is well known, the ISSN is attributed both to periodicals and to series and monographic series.

[16] For the alphabetical list of Italian periodicals see: https://doaj.org/search/journals?source=%7B%22query%22%3A%7B%-22filtered%22%3A%7B%22filter%22%3A%7B%22bool%22%3A%7B%22must%22%3A%5B%7B%22ter-ms%22%3A%7B%22index.country.exact%22%3A%5B%22Italy%22%5D%7D%7D%5D%7D%7D%2C%-22query%22%3A%7B%22match_all%22%3A%7B%7D%7D%7D%7D%2C%22size%22%3A50%2C%22sor-t%22%3A%5B%7B%22created_date%22%3A%7B%22order%22%3A%22desc%22%7D%7D%5D%7D

[17] http://www.depositolegale.it/journals/

[18] *Lucifero : periodico democratico-radicale. - [S. l. : s. n., 1870]- (Ancona : Tip. sociale). – volumi ; 38 cm. ((Settimanale. – Il complemento del titolo varia: periodico della Consociazione repubblicana delle Marche (1914); periodico repubblicano fondato nel 1870 (1964). – Diretto fino al 1904 da Domenico Barilari. - La tipografia varia: Stabilimento tip. cooperativo (1914); Tip. Bellomo (1964). - Descrizione basata su: anno 2, n. 27 (agosto 1871). – Il formato varia: 50 cm (1964). - Copia digitale anni 1914-1918 a: http://www.14-18.it/periodici/AFM_OM_B60_FASC184. – Da anno 146, n. 1 (ott.-dic. 2016) disponibile anche in Internet a: https://www.luciferonline.it/. - TO00188040; URB0934447; IEI0163814. Dal 2016 has title: *Lucifero nuovo

JLIS.it

*Radio orario : periodico settimanale / organo ufficiale della Unione radiofonica italiana. - Anno 1, n 1 (18 gennaio 1925)-anno 2, n. 4 (23 gennaio 1926). - Roma : La poligrafica nazionale, 1925-1926. – 1 volume : ill. ; 30 cm. ((L. 1.50 il numero. - BNI 1926-904. - CUB0705457
Copia digitale a: http://www.radiocorriere.teche.rai.it/Default.aspx

*Radiorario : organo ufficiale della U.R.I., Unione radiofonica italiana : tutti i programmi italiani ed esteri della settimana. - Anno 2, n. 5 (30 gennaio 1926)-anno 5, n. 52 (22 dicembre 1929). - Milano : EIAR, 1926-1929. - 4 volumi : ill. ; 30 cm. ((Settimanale. - Il complemento del titolo cambia. – UM10014518
Autore: Ente italiano audizioni radiofoniche
Copia digitale a: http://www.radiocorriere.teche.rai.it/Default.aspx

*Radiocorriere : settimanale dell'EIAR. - Anno 6, n. 1 (5/11 gennaio 1930)-anno 19, n. 37 (12-18 settembre 1943). - Torino : EIAR, 1930-1943. - 14 volumi : ill. ; 42 cm. ((Il complemento del titolo varia. - Il formato varia. – TO00202876
Autore: Ente italiano audizioni radiofoniche
Soggetto: Radiotrasmissioni – Periodici
Copia digitale a: http://www.radiocorriere.teche.rai.it/Default.aspx

*Segnale radio : settimanale dell'Eiar / Ente italiano audizioni radiofoniche. - Anno 1, n. 1 (27 ago.-2 set. 1944)-anno 2, n. 17 (22-28 aprile 1945). - Torino : S.I.P.R.A., 1944-1945 (Torino : Tipografia della S.E.T.). – 2 volumi : ill. ((Direttore Cesare Rivelli. – TO00195117
Autore: Ente italiano audizioni radiofoniche
Soggetto: Radiodiffusione – Italia – Periodici
Copia digitale a: http://www.radiocorriere.teche.rai.it/Default.aspx

*Segnale radio : musica e propaganda radiofonica nell'Italia nazifascista, 1943-1945 / Gioachino Lanotte. - Perugia : Morlacchi editore U. P., 2014. - 387 p. ; 22 cm. - BNI 2015-2626. – LIA0965392
Fa parte della collezione: *Storia
Autore: Lanotte, Gioachino
Soggetto: Fascismo - Propaganda radiofonica - Ruolo [della] Musica - Italia - 1943-1945
Classe: D384.540945

*Radiocorriere / Radio audizioni Italia. - **Ed. per l'Italia centro-meridionale**. - Anno 1, n. 1 (novembre 1945)-anno 3 (1947). - Roma : Rai, Radio Audizioni Italia, 1945-1947. – 3 volumi in folio. ((Settimanale. - BNI 1949-2985. - CFI0362950

*Radiocorriere : organo ufficiale della radio italiana. - Anno 23, n. 1 (6-12 gennaio 1946)-anno 35, n. 18 (4-10 maggio 1958). - Torino : S.I.P.R.A, 1946-1958 (Torino : S.E.T.). - 13 volumi : ill. ; 42 cm. ((Il complemento del titolo varia. - Il formato varia. – TO00202876
Variante del titolo: *Radio corriere
Soggetto: Radiotrasmissioni – Periodici
Copia digitale a: http://www.radiocorriere.teche.rai.it/Default.aspx
*Radiocorriere TV. - Anno 35, n. 19 (11/17 maggio 1958)-anno 62, n. 49 (dicembre 1985). - Torino [etc.] : [Edizioni radio italiana], 1958-1985. – 28 volumi : ill. ; 35 cm. ((Settimanale. - Il formato varia. - BNI 58-9386. - RAV0024443
Copia digitale a: http://www.radiocorriere.teche.rai.it/Default.aspx

*TV radiocorriere. - Anno 62, n. 50 (dicembre 1985)-anno 72, n. 53 (31 dicembre 1995); anno 69 (1999)-   . - Roma : Nuova Eri, 1985-2008. – volumi : ill. ; 28 cm. ((Settimanale. – Direttore Willy Molco. - CFI0398854
Ha come supplemento: *Italiana [PE. 11647]
Copia digitale 1985-1995 a: http://www.radiocorriere.teche.rai.it/Default.aspx

Radiocorriere *bibliographic record on RIDI*

JLIS.it

Digitization makes it possible to integrate the collections owned by the library with the missing issues, freely available online, of other libraries.

A new season of cataloguing must be launched starting from the mass of documents placed on the web in recent years and freely available to users. It will concern the (few) publications not yet described, but above all it, will make significant some descriptions already present in the online catalogs.[19]

The availability of digital reproductions, especially of old publications, makes it possible to reconstruct the evolution of the titles of a publication correctly, recording the titles and consistencies of the various series.

To obtain a more accurate bibliographic record, it is sometimes necessary to unify information scattered across multiple descriptions. Some information can be derived directly from digitized copies. The digital copy of the original can give us back a record uncontaminated by the use of later printed reproductions.

The comparison of different editions of digital copies can reveal or confirm the presence of parallel publications, not detected in the historical cataloguing, even of important journals.[20]

In some cases, it will be necessary to establish connections that were non-existent in the catalogs and to create descriptions with the correct serial nature.[21]

Through the analytic cataloguing and filing of the various digital libraries, it is possible to reconstruct a more complete history of the publications, starting from the observation and comparison of different issues of the same publication that may be in different cases and not communicating

---

[19] For example, going from a description like this: Il *consigliatore : giornale politico, istruttivo, letterario e commerciale. - Pinerolo, 1849-1850. - - TO00182029 to a description like this: Il *consigliatore : giornale politico, istruttivo, letterario e commerciale. – Anno 1 (1849)-anno 2, n. 18 (22 febbraio 1850). – Pinerolo : Tipografia Lobetti-Bodoni, 1849-1850. – 18 volumi. ((Settimanale. – Poi: giornale della città e provincia di Pinerolo. - Direttore: Lorenzo Giribaldi. – Descrizione basata su: Anno 1, n. 3 (10 novembre 1849). - TO00182029. Copia digitale a: https://www.giornalidelpiemonte.it/edizionitesta.php?testata=Consigliatore

[20] La *voce. - Edizione politica. - Anno 7, n. 1 (7 maggio 1915)-anno 7, n. 14 (dicembre 1915). - Roma : Libreria della Voce, 1915. – 14 volumi ; 26 cm. ((Bimensile. – Direttori: Giuseppe Prezzolini; poi: A. De Viti De Marco. – Copertina di colore giallo. - Copia digitale a: https://fondazionefeltrinelli.it/fonte/la-voce-edizione-politica-1915/#top. - TO00197733 Variante del titolo: La *voce. Edizione politica. Autore: Prezzolini, Giuseppe

[21] From the digital library of INEA, the National Institute of Agricultural Economics, we have developed this example: L'*annata agricola … nel Veneto : prime valutazioni / Andrea Povellato. – 1988-2000. - Padova : Osservatorio di economia agraria per il Veneto ed il Trentino Alto Adige, 1989-2001. – 13 volumi ; 24 cm. ((Annuale. - Poi: INEA, Istituto nazionale di economia agraria, Osservatorio di economia agraria per il Veneto. – I curatori variano. - CFI0521760. Fa parte di: *Pubblicazioni a cura dell'Osservatorio di Economia Agraria per il Veneto. Autore: Bortolozzo, Davide; Cesaro, Luca; Gambarin, Luigi; INEA; Kuehl, Gerhard; Osservatorio di politica agraria per il Veneto; Povellato, Andrea; · Schiavon, Stefano <1971->. Copia digitale: -1994-1998, 2000 a: http://dspace.crea.gov.it/handle/inea/1032/browse?type=dateissued&submit_browse=Data+di+pubblicazione -1999 a: http://dspace.crea.gov.it/bitstream/inea/1269/1/VEN-19.pdf
L'*andamento del settore agroalimentare nel Veneto : prime valutazioni per il … / [Andrea Povellato, Stefano Schiavon, Mauro Capriotti, Filippo Codato]. – 2001-2002. - Legnaro (Pd) : Veneto Agricoltura, 2002-2003. – 2 volumi : ill. ; 24 cm. ((Annuale. – Sul frontespizio: Veneto Agricoltura, in collaborazione con Inea. - PUV0880096; PUV0946606. Autore: Povellato, Andrea. Disponibile anche in Internet a: http://dspace.crea.gov.it/handle/inea/1235/browse?type=dateissued&submit_browse=Data+di+pubblicazione. *Prime valutazioni … sull'andamento del settore agroalimentare Veneto / Veneto Agricoltura ; in collaborazione con INEA, Osservatoria economico per il sistema agroalimentare e lo sviluppo rurale. – 2003-2008. - Legnaro PD : Veneto Agricoltura, 2004-2009. – 6 volumi ; 24 cm. ((Annuale. Autore: INEA; Veneto agricoltura. Disponibile anche in Internet a: http://dspace.crea.gov.it/handle/inea/904/browse?type=dateissued&submit_browse=Data+di+pubblicazione

JLIS.it

with each other.[22] We may digitize supplements, without reference to the mother journal, of which we don't even know the bibliographical description.

By comparing digitizations with catalogs and other repertories on the periodical press, we can better define the number of digitized issues compared to those published.[23] The matching between images and bibliographic descriptions must be precise, otherwise we risk keeping publications that are usable hidden. Error is always around the corner, especially in the case of publications with the same title and from the same period.

Through digital copies, we can correct erroneous information in the catalog, related to numbering and possible relationships with homogeneous periodicals. The bibliographical investigation allows us to uncover anomalies in periodicals issues and particular numbering systems.

The comparison of digitizations and previous bibliographical descriptions allows us to determine the periodicity of publications. From the reading of editorials, we can also detect cessations of periodicals and changes of titles.

Assessment elements for a quality digital library

The survey of the 73 digital libraries visited this year has allowed us to define what should be the quality criteria for a national digital library. A ranking was compiled that identifies fourteen criteria:

1. display
2. graphics
3. quality of the alphabetical sorting by titles
4. simplicity, speed, and effectiveness of the search
5. presence (or not) of a bibliographic description of the digitized material
6. presence (or not) of a bibliographic history of the publication
7. linking between the various titles of the publication
8. information about digitized holdings
9. accuracy and precision of the information
10. information about the number of digitized volumes
11. quality of the image display system
12. quality of the images
13. rarity and value of the collections
14. completeness of the digitized collections

---

[22] To reconstruct the history of The Worker of Trieste we consulted: the SBN OPAC, Internet culturale website, Biblioteca Attilio Hortis of Trieste website, Stampa clandestina website, Wikipedia and Archivio della Federazione di Trieste del Partito della Rifondazione comunista:
http://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8331
http://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8332
http://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8335
http://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8336
http://www.stampaclandestina.it/?page_id=116&ricerca=253
http://www.rifondazionecomunistatrieste.org/archivio.htm
[23] *La *guerra : pubblicazione settimanale, illustrata.* - Anno 1, n. 1 (27 giugno 1915)-n. 13 (1915). - Roma : Quattrini, 1915. – 1 volume : ill. ; 36 cm. ((BNI 1915-7778. – Copia digitale dei n. 1-10 a: http://www.14-18.it/periodici/CFI0355788/1915. - CFI0355788. Soggetto: Guerra mondiale 1914-1918.

JLIS.it

The Digital Newspaper Library of BNCR

The Digital Newspaper Library of the National Central Library of Rome (BNCR) will ideally host the bibliographic record and be identified by an alphanumeric code.[24]

Since it participated in the first European digitization projects, the BNCR has started a constant process of digitization of its collections, increased with materials coming from the participation in European projects and the collaboration with other libraries, organizations, Italian and international institutions. Among the main ones, we recall the Europeana 14-18 Project,[25] which has provided for the digitization of 20,000 images of periodicals and historical newspapers; the GoogleBooks Project which,[26] under the coordination of BNCR for Italy, has led to the digitization of over 60,000 volumes of periodicals from the period between 1668 and 1946, merged in the collections of the Digital Newspaper Library; a five-year agreement, signed in 2017 between the National Library and the Library of the Senate of the Republic "Giovanni Spadolini",[27] on the implementation of the National Newspaper Library as a single portal of access to the digitized collections of historical newspapers and journals belonging to the two libraries.

With its 2,230 titles of newspapers, periodicals, and historical journals and a patrimony of over 18 million images, the BNCR Digital Newspaper Library represents one of the richest digital newspaper libraries available on the Italian scene, continuing a long historical tradition that has involved the Biblioteca Nazionale centrale di Roma since 1908 with the task of establishing and preserving the National Newspaper Library. [28]

The available titles are based on METS for the encoding of all descriptive, administrative and structural metadata for the management of digital objects. The creation of an intermediate level, between the list of titles and the list of available years, containing a tab for each record would allow the management of bibliographic and technical information according to a Dublin Core schema.

---

[24] http://digitale.bnc.roma.sbn.it/tecadigitale/emeroteca/classic

[25] http://www.14-18.it/

[26] http://www.bncrm.beniculturali.it/it/832/progetto-googlebooks.

[27] http://digitale.bnc.roma.sbn.it/tecadigitale/progettoConvenzioneBS

[28] Andrea De Pasquale, *Per un'emeroteca nazionale digitale*, «Bibliothecae.it», 7 (2018), n. 2: 348-370, <https://bibliothecae.unibo.it/article/view/8951>.

JLIS.it

Proposed structure for a national digital newspaper library

The structure that we imagine has as a basis a super-record of the Work[29] marked by a unique and univocal alphanumeric code, on the model of Wikipedia. It is necessary to avoid the proliferation of descriptions for the same publication.

SEARCH MASKS (Access to Work)

FIRST SEARCH MASK
1. Search by Title. Browse a list of titles
2. Search by author. Browse a list of authors (Authority file)
3. Search by subject. Browse a list of subjects (Thesaurus)

The lists are sorted alphabetically and asyndetically, i.e. by significant word excluding articles and also conjunctions and prepositions if they are not at the beginning of the title. The lists can be divided into 26 blocks corresponding to the letters of the alphabet.

Search by title	Search by author	Search by subject
Antologia	Gabinetto scientifico letterario G. P. Vieusseux	Arte
Nuova antologia	Protonotari, Francesco	Cultura
Nuova antologia di lettere, scienze ed arti	Spadolini, Giovanni	Letteratura
Nuova antologia di Scienze lettere ed arti	Vieusseux, Giovan Pietro	Scienze

Example 1. Search channels

The 12 search channels are all connected to the super-record that we will call **IT2**.
The elements that the super-record should contain are:
   a) Bibliographical description
   b) Digitized volumes with links to the digital libraries
   c) Historical and bibliographical information
   d) Notes and bibliographical references
   e) Technical notes on digitization

---

[29] See IFLA, *Functional requirements for bibliographic records. Final report*, 1998.

JLIS.it

A. Bibliographical description

Antologia. - Tomo 1, n. 1 (gennaio 1821)-vol. 48, n. 144 (dicembre 1832). - Firenze : al Gabinetto scientifico e letterario di G. P. Vieusseux, 1821-1832. - 48 volumi ; 22 cm. ((Mensile. - Dal 1831 ha il complemento del titolo: giornale di scienze, lettere ed arti. – Disponibile anche in Internet come banca dati e copia digitale a: http://www.antologia-vieusseux.org/. - ISSN 1125-3622. - LO10020689

Autore: Gabinetto scientifico letterario G. P. Vieusseux

Soggetti: Arte – Periodici; Letteratura – Periodici; Scienze - Periodici

Indice generale alfabetico delle materie contenute nell'Antologia, giornale fiorentino diretto da Gio. Pietro Vieusseux : 1821-1832. - Firenze : A. Cecchi, 1863. - 270 p. ; 23 cm. - CFI0557156

Nuova antologia di scienze, lettere ed arti. - Vol. 1, fasc 1 (31 gennaio 1866)-vol. 30, fasc. 12 (dicembre 1875); 2. serie, vol. 1, fasc. 1 (gennaio 1876)-vol. 54, fasc. 24 (16 dicembre 1885); 3. serie, vol. 55, fasc. 1 (1 gennaio 1886)-vol. 60, fasc. 24 (15 dicembre 1895); 4. serie, vol. 61, fasc. 1 (1 gennaio 1896)-vol. 84, fasc. 672 (16 dicembre 1899). - Firenze : Direzione della Nuova antologia, 1866-1899. - 84 volumi : ill. ; 24 cm. ((Mensile; bimensile (1878-1880). – Fondata da Francesco Protonotari. - Dal 1876 fasc. hanno doppia numerazione. - L'editore varia. - Indici 1866-1895. – ISSN 1125-3630. - LO10020526

Nuova antologia di scienze, lettere ed arti : indice generale dei 30 volumi della prima serie : anni 1866-1875. - Firenze : Direzione della Nuova antologia, 1876. - IV, 128 p. ; 24 cm. - TSA0336581

Nuova antologia di lettere, scienze ed arti. - 4. ser., vol. 85, fasc. 673 (1 gen. 1900)-vol. 120, fasc. 816 (16 dic. 1905); 5. ser., vol. 121, fasc. 817 (1 gen. 1906)-vol. 180, fasc. 1054 (16 dic. 1915); 6. ser., vol. 181, fasc. 1055 (1 gen. 1916)-vol. 244, fasc. 1290 (16 dic. 1925); 7. ser., vol. 245, fasc. 1291 (1 gen. 1926)-vol. 246, fasc. 1298 (21 apr. 1926). - Roma : Nuova antologia, 1900-1926. - 160 volumi : ill. ; 26 cm. ((Quindicinale. - Doppia numerazione dei volumi. - Numeraz. dei fasc. progressiva negli anni. - Il vol. 234 errato nella doppia numerazione. - ISSN 1125-3649. - RAV0105511

Nuova antologia : rivista di lettere, scienze ed arti. - 7. serie, anno 61, vol. 247, fasc. 1299 (1 maggio 1926)-   . - Roma : Nuova antologia, 1926-   . - volumi ; 24 cm. ((Quindicinale; la periodicità varia. - Dal fasc. 2125/2126 (gen.-giu. 1978) il sottotitolo varia in: rivista trimestrale di lettere, scienze ed arti / diretta da Giovanni Spadolini. – Il luogo e l'editore variano in: Firenze : Le Monnier. - Indici: 1866-1985. – Copia digitale 1926-1940 a: http://digitale.bnc.roma.sbn.it/tecadigitale/giornali/RAV0027419. -RAV0027419

Soggetti: Cultura - Periodici

Classe: D055.1

Indici per autori e per materie della Nuova antologia : dal 1931 al 1950 / compilati da Laura Giuliani. - RMS0049318

Indici per autori e per materie della Nuova antologia : dal 1866 al 1930 / a cura di Lodovico Barbi. - Rist. anast. - XXIII, 721 p. ; 24 cm.

Indici 1866-2003 Disponibili in Internet all'indirizzo: https://nuovaantologia.it/storia-nuova-antologia/testi-in-pdf/

B. Digitized volumes with links to the digital libraries

Antologia 1821-1832: http://www.antologia-vieusseux.org/

Antologia 1821-1832: http://www.internetculturale.it/it/16/search?q=&searchType=avanzato&channel__creator=%-22Gabinetto+scientifico+letterario+G.+P.+Vieusseux%22&channel__contributor=%22Gabinetto+scientifico+letterario+G.+P.+Vieusseux%22&opCha__contributor=OR&opCha__creator=OR&__meta_typeLivello=periodico&pag=1

Antologia 1821-1822; 1826-1832: http://digitale.bnc.roma.sbn.it/tecadigitale/giornali/LO10020689

Nuova antologia 1926-1940: http://digitale.bnc.roma.sbn.it/tecadigitale/giornali/RAV0027419

Indici 1866-2003: https://nuovaantologia.it/storia-nuova-antologia/testi-in-pdf/

C. Historical and bibliographical information

Antologia fu una rivista con periodicità mensile, pubblicata a Firenze dal 1821 al 1833, promossa da Giovan Pietro Vieusseux e da Gino Capponi, cui collaborarono molti intellettuali del tempo.

L'indirizzo della rivista fu sempre nazionale, intendendo abbracciare i problemi generali della cultura italiana del periodo. Prima di dar vita alla rivista, Vieusseux aveva istituito, con sede a palazzo Buondelmonti, un "gabinetto scientifico-letterario" (il celebre Gabinetto Vieusseux) che, oltre a far conoscere la stampa italiana e straniera, diventò un luogo di incontri e discussioni. Furono collaboratori dell'*Antologia* quasi tutti gli intellettuali attivi fra il 1821 e il 1831, tra i quali Giuseppe Poerio, Gabriele Pepe, Pietro Colletta, Pietro Giordani, Niccolò Tommaseo, Giuseppe Montanelli, Francesco Domenico Guerrazzi, Carlo Cattaneo e Giuseppe Montani. Vieusseux fu il primo editore che compensò i propri collaboratori. Fino ad allora infatti, in Italia le collaborazioni non venivano retribuite.

JLIS.it

Pur accogliendo le istanze più disparate, la rivista vantava un orientamento comune: una preoccupazione pedagogica, che si sviluppava in chiave antirivoluzionaria; una filosofia eclettica, che escludeva però le ideologie radicali dell'Illuminismo; un'idea di "letteratura impegnata" per fini utili. Sulla rivista le questioni letterarie ebbero un posto marginale, mentre ci si occupò sistematicamente di argomenti sociali (storia, diritto, ecc.) ed economici (economia, statistica, ecc.).

Sul numero di novembre-dicembre 1832 due articoli incontrarono i rigori della censura preventiva, uno dei quali conteneva critiche all'Austria. L'uscita fu ritardata al gennaio 1833. Le autorità chiesero al direttore di rivelare i nomi degli autori dei due pezzi. Al rifiuto del direttore di uniformarsi alla decisione governativa, la rivista fu chiusura d'autorità da parte del granduca Leopoldo II di Toscana, su pressione dell'Austria.

L'*Antologia* fu per una decina di anni un elemento centrale della cultura italiana, superando di gran lunga, coi suoi oltre 500 abbonati, il numero di lettori delle riviste milanesi (si pensi al Conciliatore): la diffusione delle idee della rivista promosse la nascita di una borghesia liberale in Toscana e contribuì alla formazione del concetto di egemonia culturale

D. Notes and bibliographical references

*Paolo Prunas, *L' «Antologia» di Gian Pietro Vieusseux. Storia di una rivista italiana*, Roma, Società editrice Dante Alighieri, 1906

Antologia della «Antologia» (1821-1832). Rassegna di una rivista, a cura di Emiliano Zazo, 2 voll., Milano, Bompiani, 1945

*Umberto Carpi, *Letteratura e società nella Toscana del Risorgimento. Gli intellettuali dell'«Antologia»*, Bari, De Donato, 1974

*Angiola Ferraris, *Letteratura e impegno civile nell'«Antologia»*, Padova, Liviana, 1978

E. Technical notes on digitazion

La digitalizzazione della Biblioteca nazionale centrale di Roma è tratta da microfilm.

La digitalizzazione del Gabinetto Viesseux è iniziata nel 2015.

IT2 super-record

Title	Publication place	Publication date	Author	Subject	Bibliographic record code
Antologia	Firenze	1821-1830	Gabinetto scientifico letterario G. P. Vieusseux	Arte	IT2
Nuova antologia	Roma	1926-	Protonotari, Francesco	Cultura	IT2
Nuova antologia di lettere, scienze ed arti	Roma	1900-1926	Spadolini, Giovanni	Letteratura	IT2
Nuova antologia di scienze lettere ed arti	Firenze	1866-1899	Vieusseux, Giovan Pietro	Scienze	IT2

Second search mask (Database)

The search for individual issues in the digital libraries, especially for newspapers and weeklies that may include thousands of units, should not proceed by overall chronological browsing, but broken down into years, months and days, possibly using predefined chronological grids that make it easier to locate the issues sought.

Example 1: https://www.giornalidelpiemonte.it/edizionitesta.php?testata=Il%20Biellese
Il biellese, a biweekly with 1145 pages of a search for individual issues. The chronological browsing is annoying, also because for each search the system brings back to the initial page, and therefore to search for a month of the magazine, it is necessary to search about ten times and each time to browse all the pages of the site.

Example 2: https://avanti.senato.it/avanti/
Avanti! from the Senate Library. With just a few steps you get directly to the day you are looking for. It is possible to browse through the header, visualize the list of the digitized years, select the desired year, choose the edition and the month, visualize the first pages of each day of the month with the date highlighted, once a search is carried out it returns to the previous screen.

Example 2. Issue search within the digital libraries

Closing remarks

In the current scenario, after more than 20 years, it is essential to rethink the cultural policies in the digital field, pooling projects, ideas, financial and human resources, overcoming inappropriate attitudes of personal or institutional pride to start building a common path for the development and use of the Italian digital heritage. We need to rediscover the united effort among public and private bodies and institutions that characterized the success of SBN in the 1980s. For the knowledge, diffusion, and valorization of the Italian digital heritage, we need a tool that resembles what SBN represents today for the bibliographic heritage of Italian libraries.

JLIS.it

References

De Pasquale, Andrea. 2018. "Per un'emeroteca nazionale digitale." *Bibliothecae.it*, 7 (2018), n. 2: 348-370. DOI 10.6092/issn.2283-9364/8951

Formenti, Carlo. *Corriere della sera*, 20 novembre 2000, 27.

Guerrini, Mauro. 1999. "Catalogare le risorse elettroniche: lo standard ISBD(ER)." *Biblioteche oggi*, 17 (1999), n. 1: 62.

IFLA. 1998. *Functional requirements for bibliographic records. Final report.* Munchen: K.G. Saur. https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

Mangiaracina, Silvana, and Morroni Cristina. 2018. *Quanto costa l'accesso alle pubblicazioni scientifiche nell'era dell'Open Access?: una prima analisi delle pubblicazioni nel CNR.* Bologna: Biblioteca Area della ricerca di Bologna CNR. https://zenodo.org/record/1247497#.XoC-JKPOPkU