JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12794

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12794

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12738

JLIS.it

Universal Bibliographic Control today: preliminary remarks

Mathilde Koskas[(a)]

a) Bibliothèque nationale de France

Contact: Mathilde Koskas, mathilde.koskas@lilo.org

Received: 15 April 2021; **Accepted:** 14 May 2021; **First Published:** 15 January 2022

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12738

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

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".

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.

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.

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 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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12782

JLIS.it

Conference BC 2021

Josep Torn[a]

a) European University Institute

Contact: Josep Torn, pep.torn@eui.eu
Received: 27 September 2021; **Accepted:** 30 September 2021; **First Published:** 15 January 2022

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.

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.

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' 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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12742

JLIS.it

Universal bibliographic control in the digital ecosystem: opportunities and challenges

Mauro Guerrini[(a)]

a) Università degli Studi di Firenze

Contact: Mauro Guerrini, mauro.guerrini@unifi.it

Received: 16 April 2021; **Accepted:** 26 August 2021; **First Published:** 15 January 2022

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12742

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/.

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).

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.

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.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12783

JLIS.it

Standards in a new bibliographic world

Renate Behrens[a]

a) German National Library

Contact: Renate Behrens, r.behrens@dnb.de

Received: 28 September 2021; **Accepted:** 30 September 2021; **First Published:** 15 January 2022

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.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12783

tool in libraries and are labour-intensively administered there – within the Integrated Authority 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 responsible 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 factored 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 generally 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 players. 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 described in the next section.

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

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.

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/.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12739

JLIS.it

Bibliographic control in the fifth information age

Gordon Dunsire[a]

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

Contact: Gordon Dunsire, gordon@gordondunsire.com

Received: 15 April 2021; **Accepted:** 12 May 2021; **First Published:** 15 January 2022

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12739

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 necessary 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.

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.

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.

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.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12739

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.

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.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12762

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

Contact: Richard Wallis, richard.wallis@dataliberate.com

Received: 7 June 2021; **Accepted:** 7 June 2021; **First Published:** 15 January 2022

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.

KEYWORDS

Bibliographic data; BIBFRAME; Schema.org.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12762

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/

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/

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?...

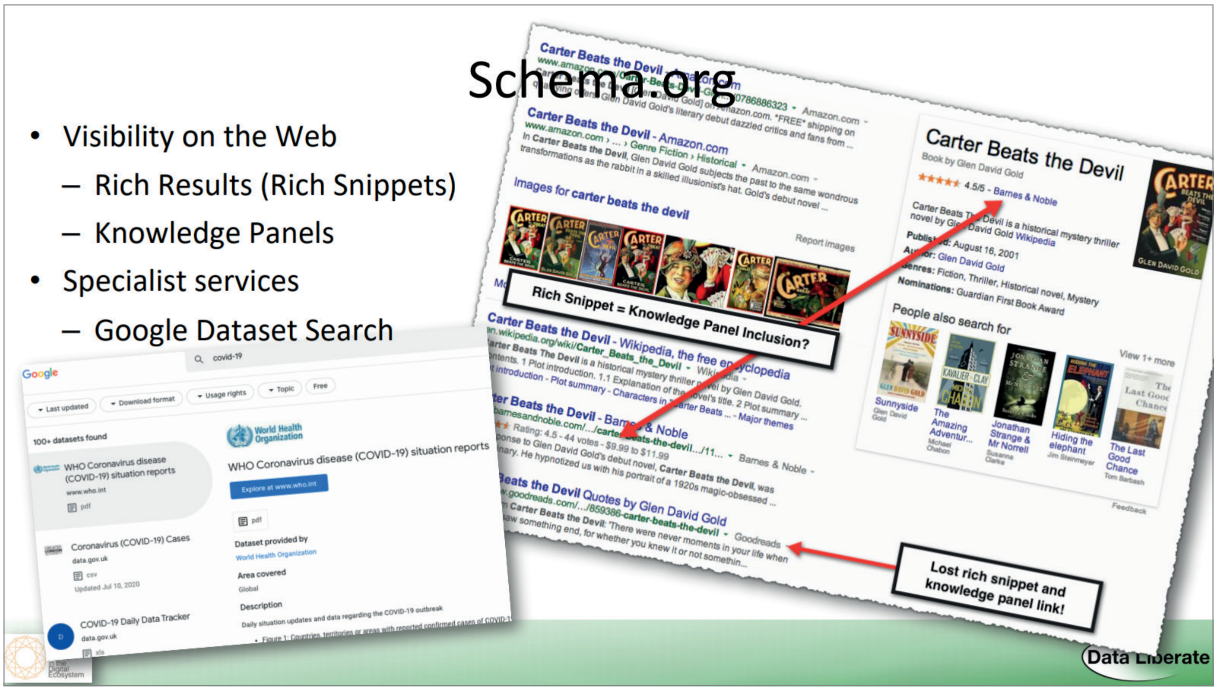
[4] http://webdatacommons.org/structureddata/2020-12/stats/stats.html
[5] https://en.wikipedia.org/wiki/JSON-LD

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12762

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.



Schema.org

- Visibility on the Web
  - Rich Results (Rich Snippets)
  - Knowledge Panels
- Specialist services
  - Google Dataset Search

Carter Beats the Devil - Amazon.com
www.amazon.com › ... › Genre Fiction › Historical · Amazon.com
In Carter Beats the Devil, Glen David Gold subjects the past to the same wondrous transformations as the rabbit in a skilled illusionist's hat. Gold's debut novel ...

Images for carter beats the devil

Rich Snippet = Knowledge Panel Inclusion?

Carter Beats the Devil - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Carter_Beats_the_Devil · Wikipedia
Carter Beats The Devil is a historical mystery thriller novel by Glen David Gold. Contents. 1 Plot introduction. 1.1 Explanation of the novel's title. 2 Plot summary ... Plot introduction - Plot summary - Characters in Carter Beats ... - Major themes

Carter Beats the Devil - Barnes & Noble
barnesandnoble.com/.../carter-beats-the-devil.../11... · Barnes & Noble
★★★★★ Rating: 4.5 - 44 votes - $9.99 to $11.99
Response to Glen David Gold's debut novel, Carter Beats the Devil, was ... He hypnotized us with his portrait of a 1920s magic-obsessed ...

Carter Beats the Devil Quotes by Glen David Gold
www.goodreads.com/.../859386-carter-beats-the-devil · Goodreads
Carter Beats the Devil: 'There were never moments in your life when you saw something end, for whether you knew it or not somethin...

Carter Beats the Devil
Book by Glen David Gold
★★★★★ 4.5/5 - Barnes & Noble
Carter Beats The Devil is a historical mystery thriller novel by Glen David Gold Wikipedia
Published: August 16, 2001
Author: Glen David Gold
Genres: Fiction, Thriller, Historical novel, Mystery
Nominations: Guardian First Book Award

People also search for
Sunnyside Glen David Gold
The Amazing Adventur... Michael Chabon
Jonathan Strange & Mr Norrell Susanna Clarke
Hiding the elephant Jim Steinmeyer
The Last Good Chance Tom Barbash

Lost rich snippet and knowledge panel link!

covid-19

100+ datasets found

WHO Coronavirus disease (COVID-19) situation reports
www.who.int
pdf

Coronavirus (COVID-19) Cases
data.gov.uk
csv
Updated Jul 10, 2020

COVID-19 Daily Data Tracker
data.gov.uk
xls

WHO Coronavirus disease (COVID-19) situation reports
Explore at www.who.int
pdf
Dataset provided by
World Health Organization
Area covered
Global
Description
Daily situation updates and data regarding the COVID-19 outbreak

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/

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.

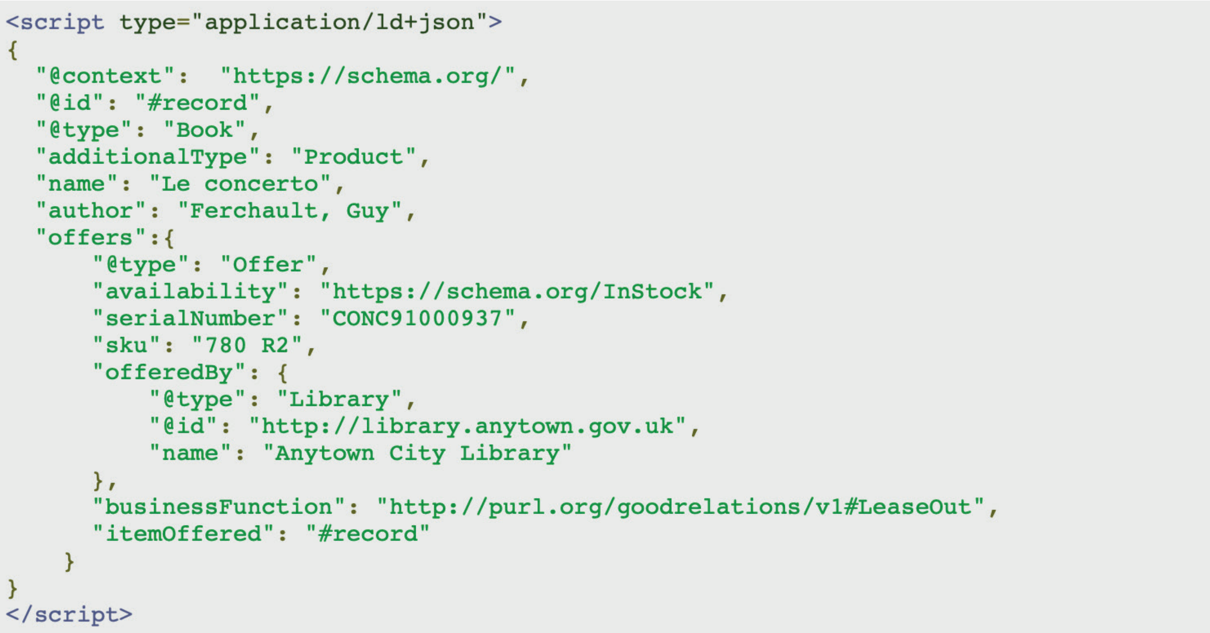
So, what are libraries doing on the web at the moment?...
Well mostly (there are some exceptions), basically the web knows about your discovery interface (hopefully) or (more likely) the homepage of your website. Not the things you can discover using that interface. Users do not start their discovery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

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 BIBFRAME 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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12760

JLIS.it

Collocation and Hubs. Fundamental and New Version

Sally H. McCallum[(a)]

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

Contact: Sally H. McCallum, smcc@loc.gov

Received: 3 June 2021; **Accepted:** 11 July 2021; **First Published:** 15 January 2022

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.

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.

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 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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12760

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.

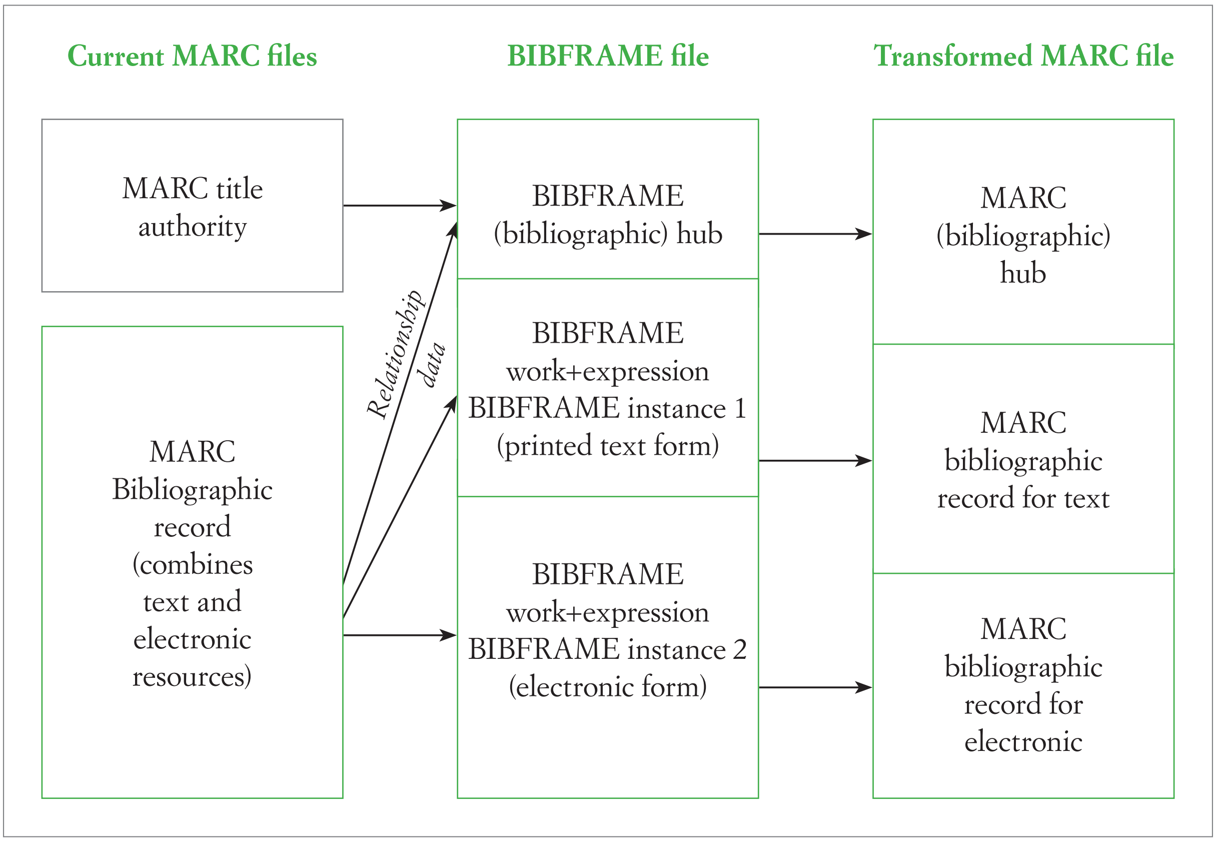

Current MARC files
BIBFRAME file
Transformed MARC file
MARC title authority
MARC Bibliographic record (combines text and electronic resources)
Relationship data
BIBFRAME (bibliographic) hub
BIBFRAME work+expression BIBFRAME instance 1 (printed text form)
BIBFRAME work+expression BIBFRAME instance 2 (electronic form)
MARC (bibliographic) hub
MARC bibliographic record for text
MARC bibliographic record for electronic

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 BIBFRAME 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 BIBFRAME 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.

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.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12748

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

Contact: Tiziana Possemato, tiziana.possemato@atcult.it

Received: 5 May 2021; **Accepted:** 23 May 2021; **First Published:** 15 January 2022

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.

KEYWORDS

Semantic Web; Real World Object (RWO); Entity reconciliation; Universal Bibliographic Control (UBC); Entity; Identity.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12748

> 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 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.

- 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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12748

- 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, 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.

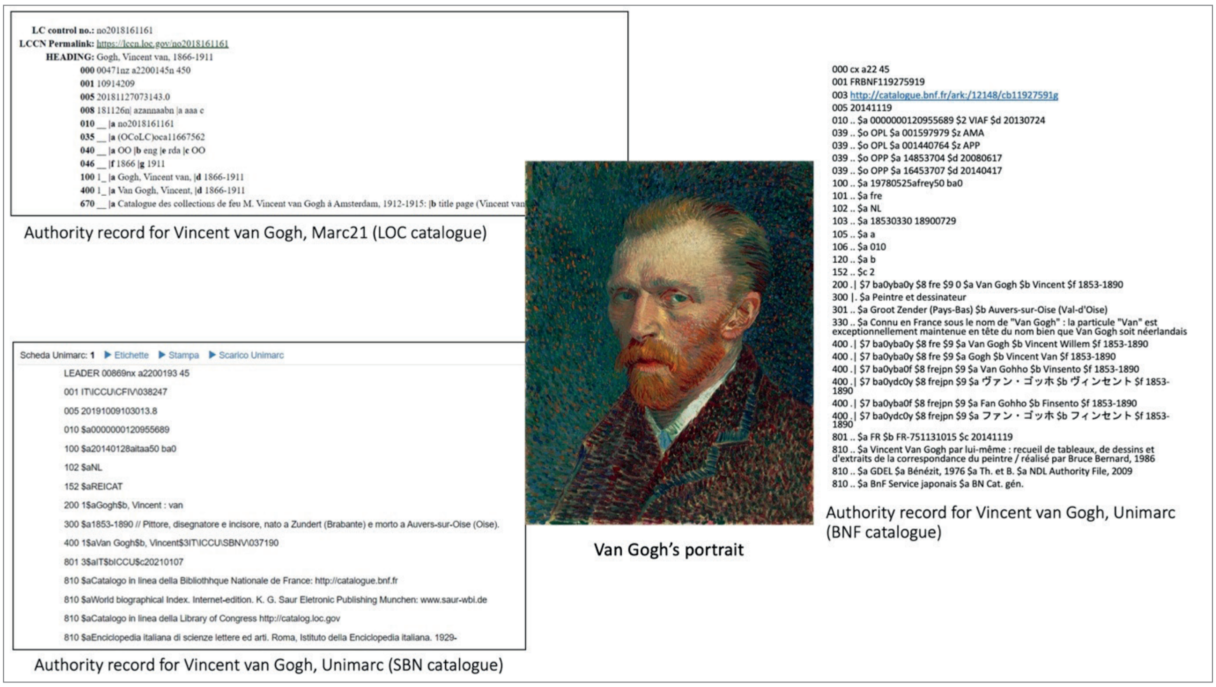
"Are the winner of Austerlitz and the loser of Waterloo the same person?"

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)>



```
LC control no.: no2018161161
LCCN Permalink: https://lccn.loc.gov/no2018161161
      HEADING: Gogh, Vincent van, 1866-1911
          000 00471nz a2200145n 450
          001 10914209
          005 20181127073143.0
          008 181126n| azannaabn |a aaa c
          010 __ |a no2018161161
          035 __ |a (OCoLC)oca11667562
          040 __ |a OO |b eng |e rda |c OO
          046 __ |f 1866 |g 1911
          100 1_ |a Gogh, Vincent van, |d 1866-1911
          400 1_ |a Van Gogh, Vincent, |d 1866-1911
          670 __ |a Catalogue des collections de feu M. Vincent van Gogh à Amsterdam, 1912-1915: |b title page (Vincent van
```

Authority record for Vincent van Gogh, Marc21 (LOC catalogue)

```
Scheda Unimarc: 1   ▶ Etichette   ▶ Stampa   ▶ Scarico Unimarc
      LEADER 00869nx a2200193 45
      001 IT\ICCU\CFIV\038247
      005 20191009103013.8
      010 $a0000000120955689
      100 $a20140128aitaa50 ba0
      102 $aNL
      152 $aREICAT
      200 1$aGogh$b, Vincent : van
      300 $a1853-1890 // Pittore, disegnatore e incisore, nato a Zundert (Brabante) e morto a Auvers-sur-Oise (Oise).
      400 1$aVan Gogh$b, Vincent$3IT\ICCU\SBNV\037190
      801 3$aIT$bICCU$c20210107
      810 $aCatalogo in linea della Bibliothhque Nationale de France: http://catalogue.bnf.fr
      810 $aWorld biographical Index. Internet-edition. K. G. Saur Eletronic Publishing Munchen: www.saur-wbi.de
      810 $aCatalogo in linea della Library of Congress http://catalog.loc.gov
      810 $aEnciclopedia italiana di scienze lettere ed arti. Roma, Istituto della Enciclopedia italiana. 1929-
```

Authority record for Vincent van Gogh, Unimarc (SBN catalogue)

Van Gogh's portrait

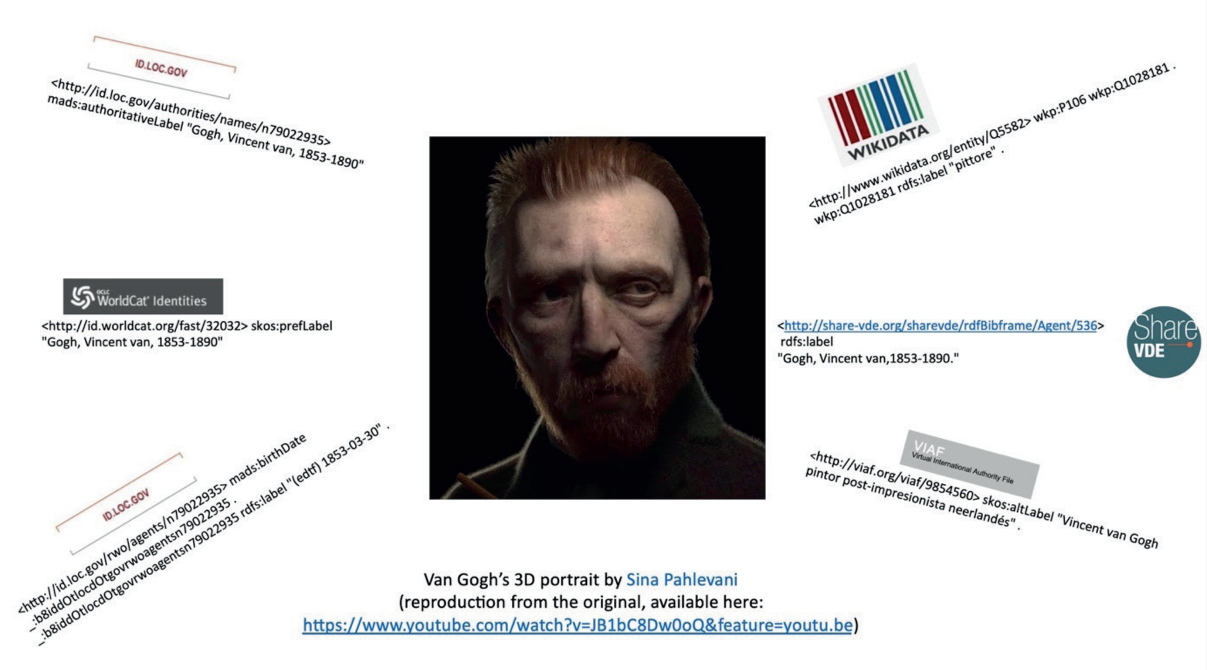
```
000 cx a22 45
001 FRBNF119275919
003 http://catalogue.bnf.fr/ark:/12148/cb11927591g
005 20141119
010 .. $a 0000000120955689 $2 VIAF $d 20130724
039 .. $o OPL $a 001597979 $z AMA
039 .. $o OPL $a 001440764 $z APP
039 .. $o OPP $a 14853704 $d 20080617
039 .. $o OPP $a 16453707 $d 20140417
100 .. $a 19780525afrey50 ba0
101 .. $a fre
102 .. $a NL
103 .. $a 18530330 18900729
105 .. $a a
106 .. $a 010
120 .. $a b
152 .. $c 2
200 .| $7 ba0yba0y $8 fre $9 0 $a Van Gogh $b Vincent $f 1853-1890
300 |. $a Peintre et dessinateur
301 .. $a Groot Zender (Pays-Bas) $b Auvers-sur-Oise (Val-d'Oise)
330 .. $a Connu en France sous le nom de "Van Gogh" : la particule "Van" est
exceptionnellement maintenue en tête du nom bien que Van Gogh soit néerlandais
400 .| $7 ba0yba0y $8 fre $9 $a Van Gogh $b Vincent Willem $f 1853-1890
400 .| $7 ba0yba0y $8 fre $9 $a Gogh $b Vincent Van $f 1853-1890
400 .| $7 ba0yba0f $8 frejpn $9 $a Van Gohho $b Vinsento $f 1853-1890
400 .| $7 ba0ydc0y $8 frejpn $9 $a ヴァン・ゴッホ $b ヴィンセント $f 1853-
1890
400 .| $7 ba0yba0f $8 frejpn $9 $a Fan Gohho $b Finsento $f 1853-1890
400 .| $7 ba0ydc0y $8 frejpn $9 $a ファン・ゴッホ $b フィンセント $f 1853-
1890
801 .. $a FR $b FR-751131015 $c 20141119
810 .. $a Vincent Van Gogh par lui-même : recueil de tableaux, de dessins et
d'extraits de la correspondance du peintre / réalisé par Bruce Bernard, 1986
810 .. $a GDEL $a Bénézit, 1976 $a Th. et B. $a NDL Authority File, 2009
810 .. $a BnF Service japonais $a BN Cat. gén.
```

Authority record for Vincent van Gogh, Unimarc (BNF catalogue)

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.



<http://id.loc.gov/authorities/names/n79022935>
mads:authoritativeLabel "Gogh, Vincent van, 1853-1890"

<http://www.wikidata.org/entity/Q5582> wkp:P106 wkp:Q1028181 .
wkp:Q1028181 rdfs:label "pittore" .

<http://id.worldcat.org/fast/32032> skos:prefLabel
"Gogh, Vincent van, 1853-1890"

<http://share-vde.org/sharevde/rdfBibframe/Agent/536>
rdfs:label
"Gogh, Vincent van,1853-1890."

<http://id.loc.gov/rwo/agents/n79022935> mads:birthDate
_:b8iddOtlocdOtgovrwoagentsn79022935 .
_:b8iddOtlocdOtgovrwoagentsn79022935 rdfs:label "(edtf) 1853-03-30" .

<http://viaf.org/viaf/9854560> skos:altLabel "Vincent van Gogh
pintor post-impresionista neerlandés" .

Van Gogh's 3D portrait by Sina Pahlevani
(reproduction from the original, available here:
https://www.youtube.com/watch?v=JB1bC8Dw0oQ&feature=youtu.be)

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?

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>

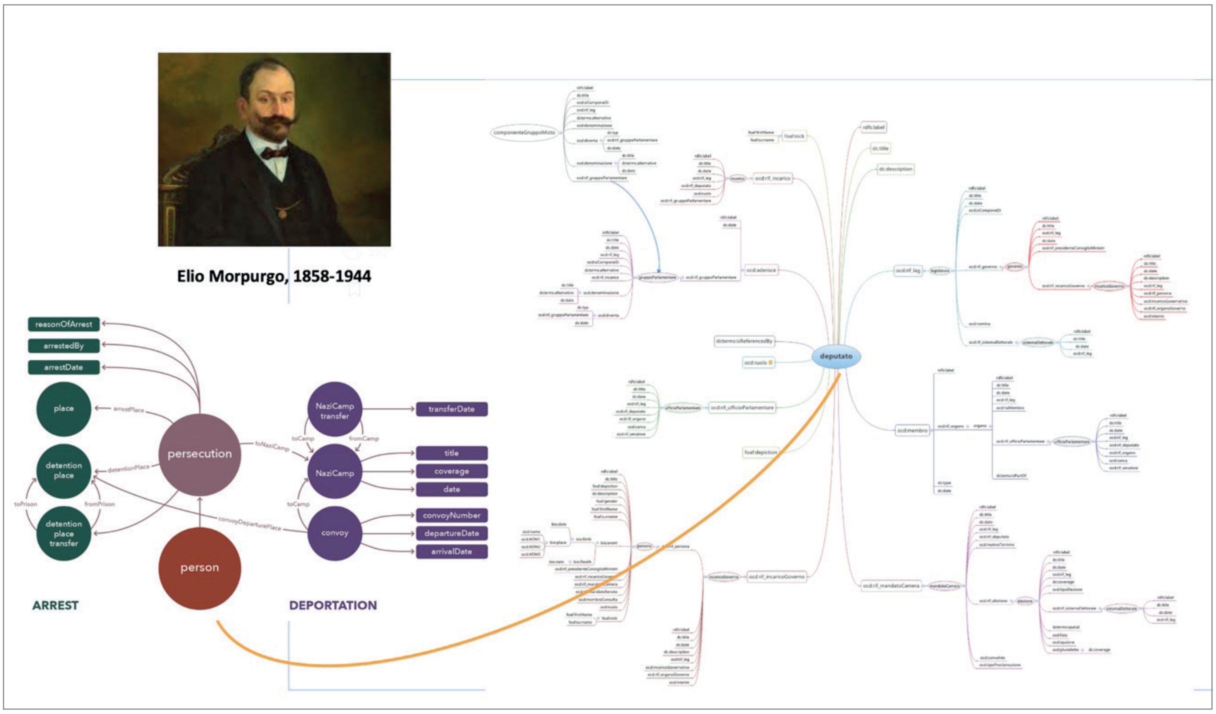


Fig. 4. The Italian politician Elio Morpurgo is identified through highly authoritative sources

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-

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12748

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.).

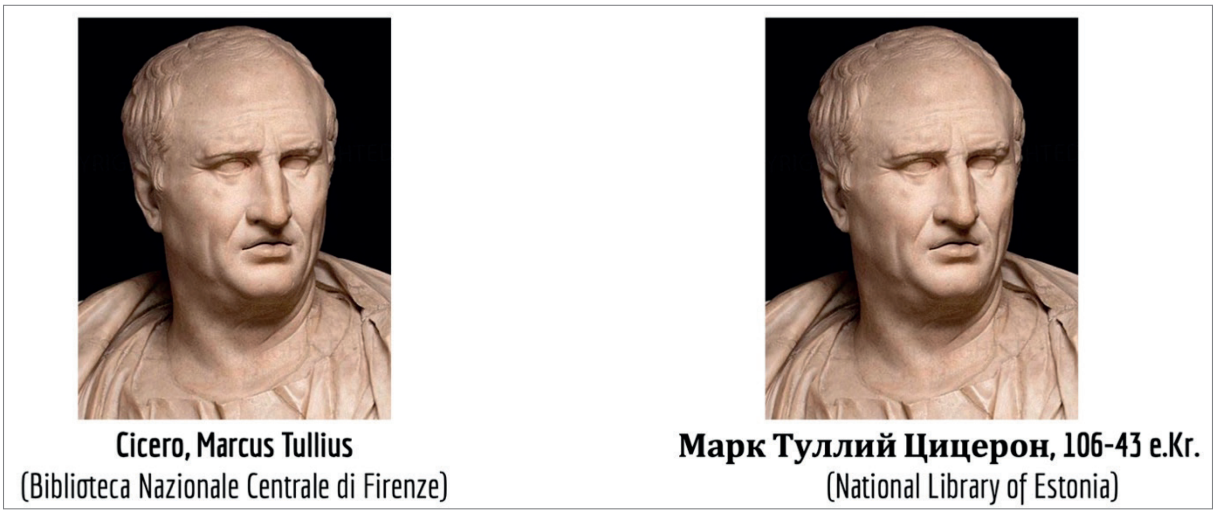

Cicero, Marcus Tullius
(Biblioteca Nazionale Centrale di Firenze)
Марк Туллий Цицерон, 106-43 e.Kr.
(National Library of Estonia)

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 – פירנדלו, לואיג’י, 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);

- 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 – פירנדלו, לואיג'י, 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.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12735

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

Contact: Ian Bigelow, bigelow@ualberta.ca; Abigail Sparling, ajsparli@ualberta.ca

Received: 14 April 2021; **Accepted:** 11 May 2021; **First Published:** 15 January 2022

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.

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.

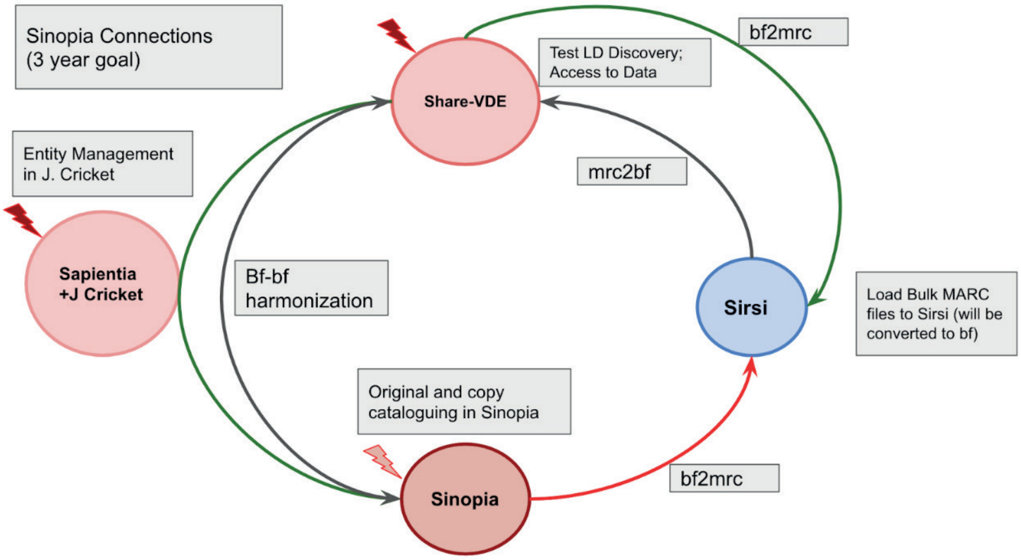


Fig. 1. Sinopia Connections (3 year goal) (Image by Bigelow, 2020)

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 projects 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).

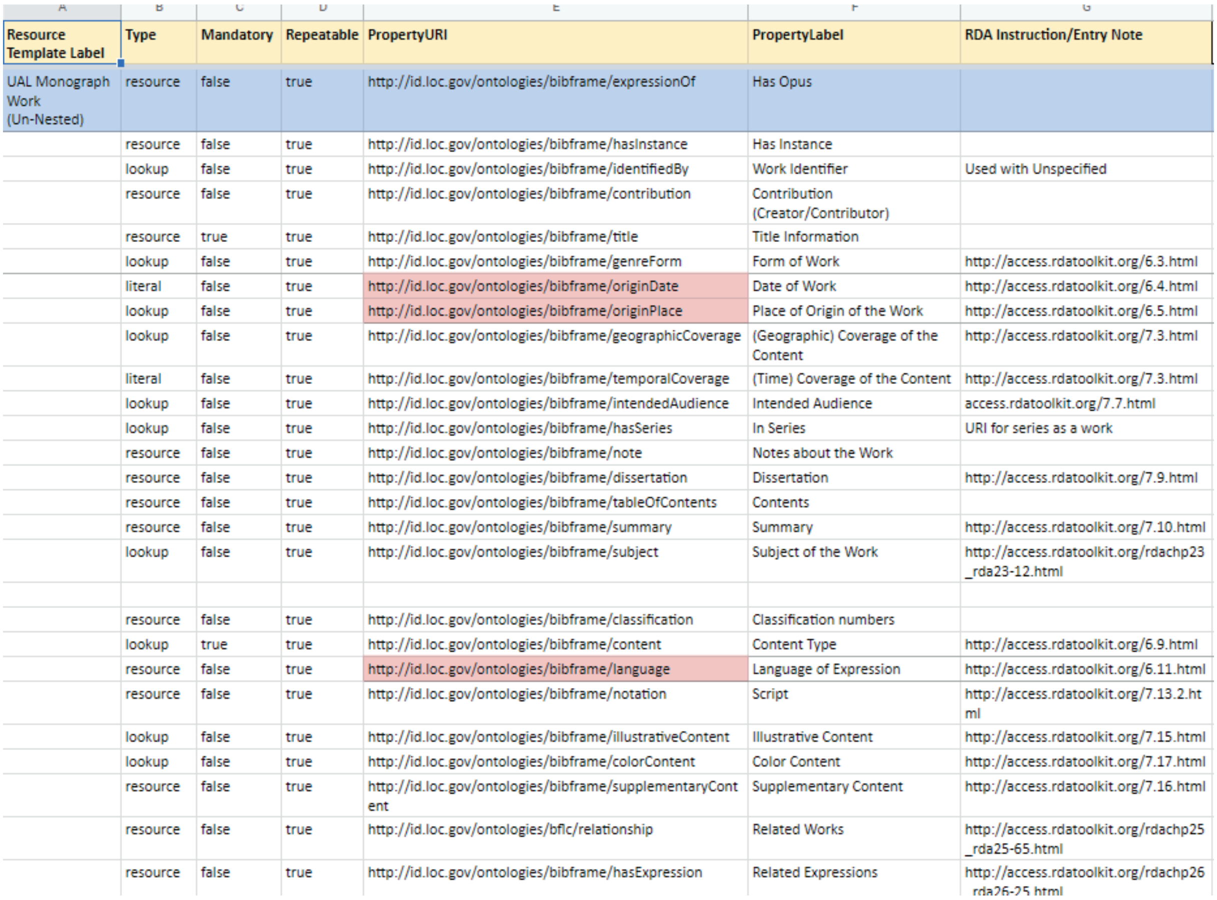
Reflecting on this timeline, 2017-2020 saw increased development not just in BIBFRAME, but in the evaluation, testing and analysis of use of RDA in a linked data environment. This acceleration has resulted in beautiful chaos, with further work on data modelling, more maturity in conversion processes, 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.

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.



A	B	C	D	E	F	G
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.

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.

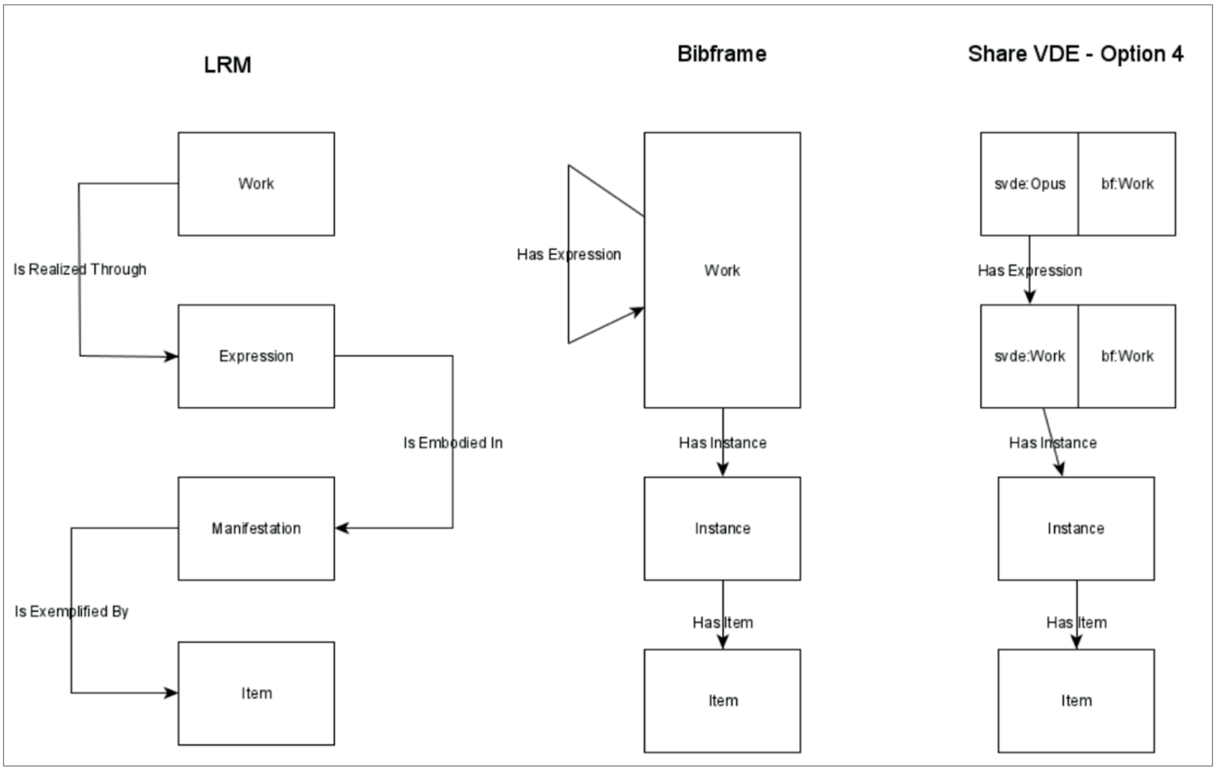


Fig. 3. LRM, BIBFRAME and Share-VDE model comparison. (Ford, Kevin 2020b. [Share-VDE - Option 4]. Created for the Share-VDE Sapientia Entity Identification Working Group)

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 highlighting 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 because [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 exciting 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).

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 BIBFRAME) 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.

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.

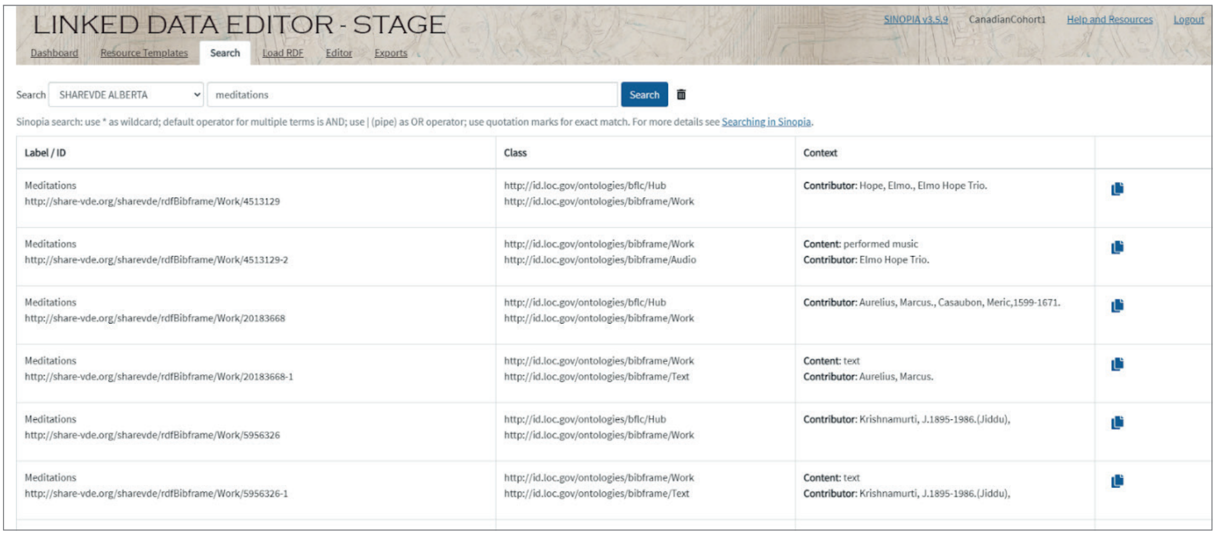
That many such variances need to be considered and decisions made for conversion, matching, and clustering again points to the desperate need for standardization, at least for core BIBFRAME elements. If these decisions are made independently for a given library or community for elements that are not 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.*

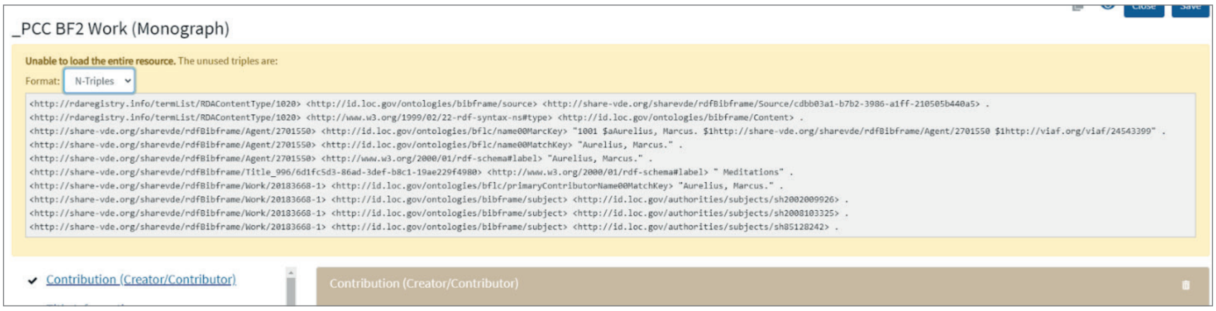


Sinopia search: use * as wildcard; default operator for multiple terms is AND; use | (pipe) as OR operator; use quotation marks for exact match. For more details see Searching in Sinopia.

Label / ID	Class	Context
Meditations http://share-vde.org/sharevde/rdfBibframe/Work/4513129	http://id.loc.gov/ontologies/bflc/Hub http://id.loc.gov/ontologies/bibframe/Work	**Contributor:** Hope, Elmo., Elmo Hope Trio.
Meditations http://share-vde.org/sharevde/rdfBibframe/Work/4513129-2	http://id.loc.gov/ontologies/bibframe/Work http://id.loc.gov/ontologies/bibframe/Audio	**Content:** performed music **Contributor:** Elmo Hope Trio.
Meditations http://share-vde.org/sharevde/rdfBibframe/Work/20183668	http://id.loc.gov/ontologies/bflc/Hub http://id.loc.gov/ontologies/bibframe/Work	**Contributor:** Aurelius, Marcus., Casaubon, Meric,1599-1671.
Meditations http://share-vde.org/sharevde/rdfBibframe/Work/20183668-1	http://id.loc.gov/ontologies/bibframe/Work http://id.loc.gov/ontologies/bibframe/Text	**Content:** text **Contributor:** Aurelius, Marcus.
Meditations http://share-vde.org/sharevde/rdfBibframe/Work/5956326	http://id.loc.gov/ontologies/bflc/Hub http://id.loc.gov/ontologies/bibframe/Work	**Contributor:** Krishnamurti, J.1895-1986.(Jiddu),
Meditations http://share-vde.org/sharevde/rdfBibframe/Work/5956326-1	http://id.loc.gov/ontologies/bibframe/Work http://id.loc.gov/ontologies/bibframe/Text	**Content:** text **Contributor:** Krishnamurti, J.1895-1986.(Jiddu),

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.


Close
Save
Contribution (Creator/Contributor)
Contribution (Creator/Contributor)

_PCC BF2 Work (Monograph)

Unable to load the entire resource. The unused triples are:

Format: N-Triples

```
<http://rdaregistry.info/termList/RDAContentType/1020> <http://id.loc.gov/ontologies/bibframe/source> <http://share-vde.org/sharevde/rdfBibframe/Source/cdbb03a1-b7b2-3986-a1ff-210505b440a5> .
<http://rdaregistry.info/termList/RDAContentType/1020> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://id.loc.gov/ontologies/bibframe/Content> .
<http://share-vde.org/sharevde/rdfBibframe/Agent/2701550> <http://id.loc.gov/ontologies/bflc/name00MarcKey> "1001 $aAurelius, Marcus. $1http://share-vde.org/sharevde/rdfBibframe/Agent/2701550 $1http://viaf.org/viaf/24543399" .
<http://share-vde.org/sharevde/rdfBibframe/Agent/2701550> <http://id.loc.gov/ontologies/bflc/name00MatchKey> "Aurelius, Marcus." .
<http://share-vde.org/sharevde/rdfBibframe/Agent/2701550> <http://www.w3.org/2000/01/rdf-schema#label> "Aurelius, Marcus." .
<http://share-vde.org/sharevde/rdfBibframe/Title_996/6d1fc5d3-86ad-3def-b8c1-19ae229f4980> <http://www.w3.org/2000/01/rdf-schema#label> " Meditations" .
<http://share-vde.org/sharevde/rdfBibframe/Work/20183668-1> <http://id.loc.gov/ontologies/bflc/primaryContributorName00MatchKey> "Aurelius, Marcus." .
<http://share-vde.org/sharevde/rdfBibframe/Work/20183668-1> <http://id.loc.gov/ontologies/bibframe/subject> <http://id.loc.gov/authorities/subjects/sh2002009926> .
<http://share-vde.org/sharevde/rdfBibframe/Work/20183668-1> <http://id.loc.gov/ontologies/bibframe/subject> <http://id.loc.gov/authorities/subjects/sh2008103325> .
<http://share-vde.org/sharevde/rdfBibframe/Work/20183668-1> <http://id.loc.gov/ontologies/bibframe/subject> <http://id.loc.gov/authorities/subjects/sh85128242> .
```

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, supported 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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12735

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.

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-BIBFRAME-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

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/1t5Kh-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/1oPEMnnpMnCIXoIMWztThm_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.

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%20IpmementationPlanNov2.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

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12737

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12737

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

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

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/

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)
```

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12737

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.

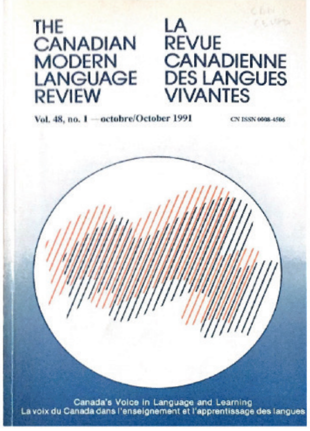
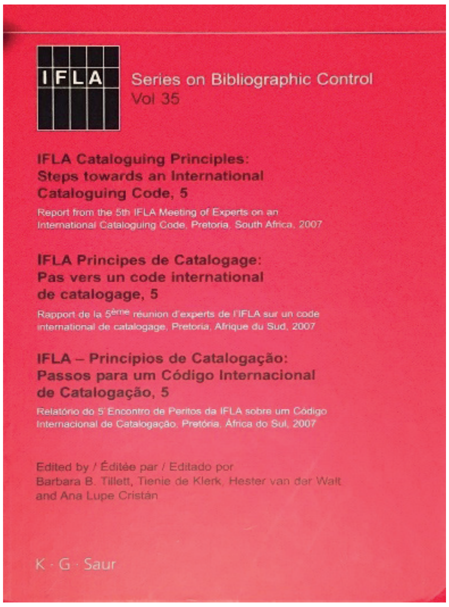

THE CANADIAN MODERN LANGUAGE REVIEW
LA REVUE CANADIENNE DES LANGUES VIVANTES
Vol. 48, no. 1 — octobre/October 1991
Canada's Voice in Language and Learning
La voix du Canada dans l'enseignement et l'apprentissage des langues

Fig. 1. Canadian Modern Languages Review

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12737



IFLA

Series on Bibliographic Control
Vol 35

IFLA Cataloguing Principles:
Steps towards an International
Cataloguing Code, 5

Report from the 5th IFLA Meeting of Experts on an
International Cataloguing Code, Pretoria, South Africa, 2007

IFLA Principes de Catalogage:
Pas vers un code international
de catalogage, 5

Rapport de la 5ème réunion d'experts de l'IFLA sur un code
international de catalogage, Pretoria, Afrique du Sud, 2007

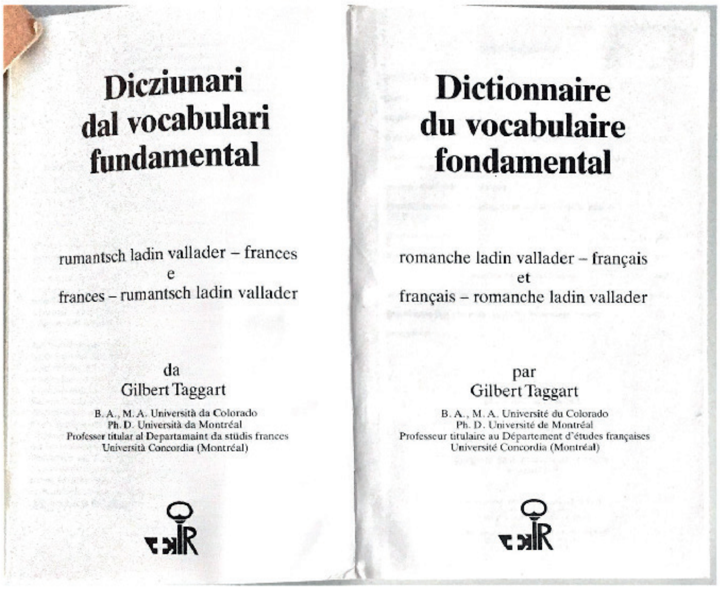
IFLA – Principios de Catalogação:
Passos para um Código Internacional
de Catalogação, 5

Relatório do 5' Encontro de Peritos da IFLA sobre um Código
Internacional de Catalogação, Pretória, África do Sul, 2007

Edited by / Éditée par / Editado por
Barbara B. Tillett, Tienie de Klerk, Hester van der Walt
and Ana Lupe Cristán

K · G · Saur

Fig. 2. IFLA Cataloguing Principles



Dicziunari
dal vocabulari
fundamental

rumantsch ladin vallader – frances
e
frances – rumantsch ladin vallader

da
Gilbert Taggart

B. A., M. A. Università da Colorado
Ph. D. Università da Montréal
Professer titular al Departamaint da stüdis frances
Università Concordia (Montréal)

Dictionnaire
du vocabulaire
fondamental

romanche ladin vallader – français
et
français – romanche ladin vallader

par
Gilbert Taggart

B. A., M. A. Université du Colorado
Ph. D. Université de Montréal
Professeur titulaire au Département d'études françaises
Université Concordia (Montréal)

Fig. 3. Rumantsch-French bilingual dictionary

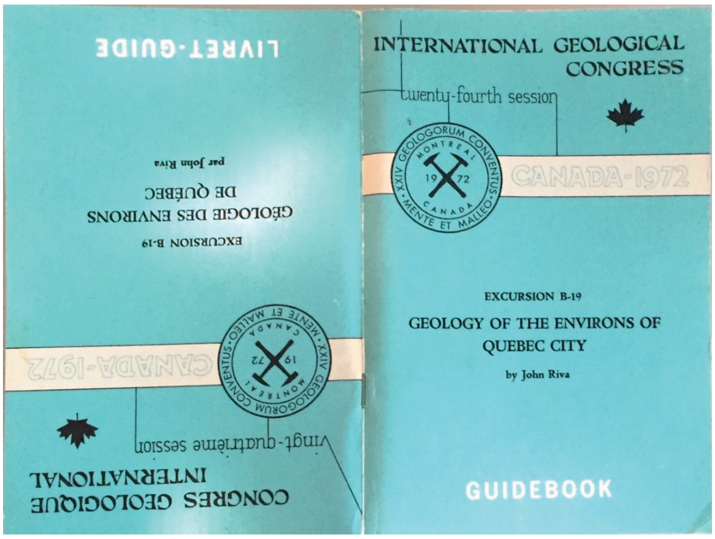
In this bilingual Rumantsch-French dictionary (figure 3) there are two full title pages. Applying the left-to-right convention again, this time to the choice of title page, the cataloguer is clearly directed to record the Rumantsch title first, as the title proper. The content of the dictionary alternates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


LIVRET-GUIDE

EXCURSION B-19
GÉOLOGIE DES ENVIRONS
DE QUÉBEC
par John Riva

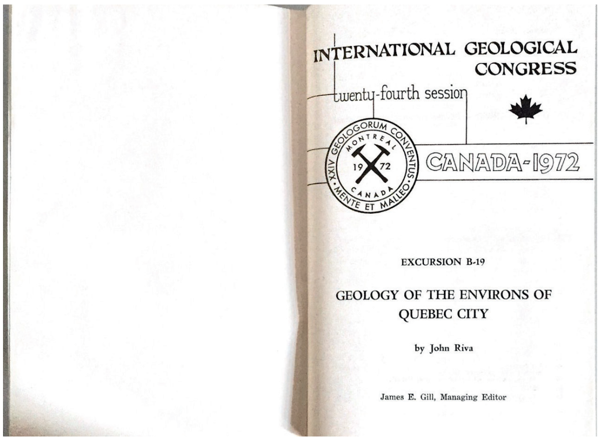
CONGRES GEOLOGIQUE
INTERNATIONAL
vingt-quatrième session
CANADA-1972

INTERNATIONAL GEOLOGICAL
CONGRESS
twenty-fourth session
CANADA-1972

EXCURSION B-19
GEOLOGY OF THE ENVIRONS OF
QUEBEC CITY
by John Riva

GUIDEBOOK

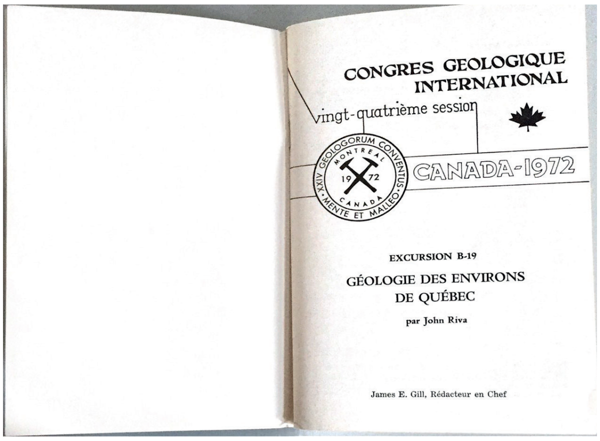
Fig. 4. Tête-bêche publication open to show both covers


INTERNATIONAL GEOLOGICAL
CONGRESS
twenty-fourth session
CANADA-1972

EXCURSION B-19
GEOLOGY OF THE ENVIRONS OF
QUEBEC CITY
by John Riva

James E. Gill, Managing Editor

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


CONGRES GEOLOGIQUE
INTERNATIONAL
vingt-quatrième session
CANADA-1972

EXCURSION B-19
GÉOLOGIE DES ENVIRONS
DE QUÉBEC
par John Riva

James E. Gill, Rédacteur en Chef

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12737

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    __ ǀa Musées
750    _0 ǀa 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.

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.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12773

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.

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.

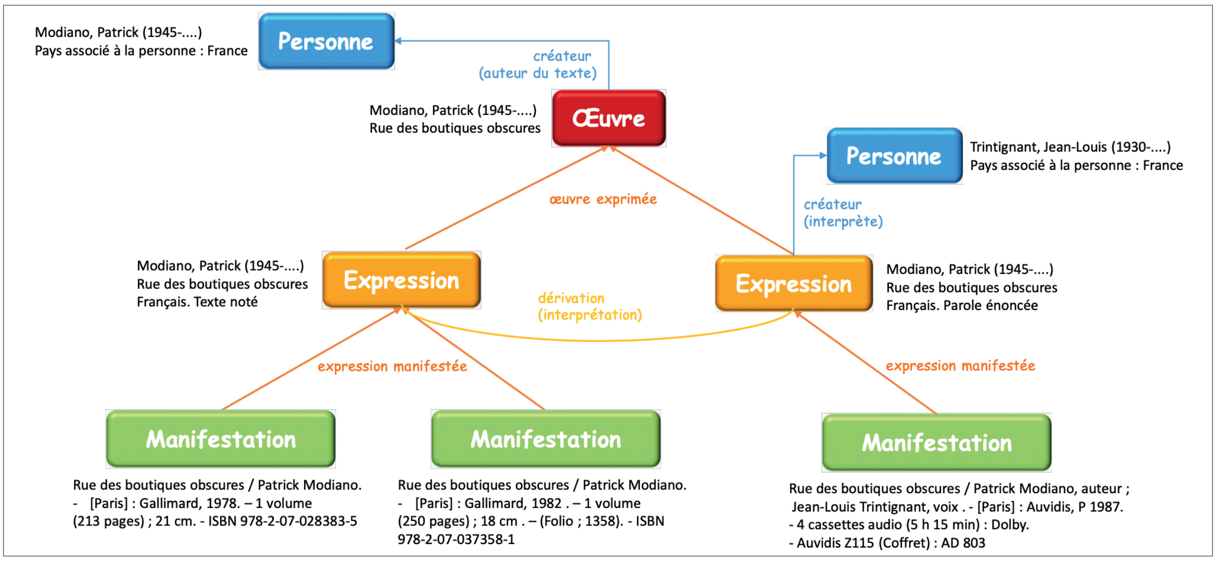

Modiano, Patrick (1945-....)
Pays associé à la personne : France
Personne
créateur
(auteur du texte)
Modiano, Patrick (1945-....)
Rue des boutiques obscures
Œuvre
Personne
Trintignant, Jean-Louis (1930-....)
Pays associé à la personne : France
œuvre exprimée
créateur
(interprète)
Modiano, Patrick (1945-....)
Rue des boutiques obscures
Français. Texte noté
Expression
dérivation
(interprétation)
Expression
Modiano, Patrick (1945-....)
Rue des boutiques obscures
Français. Parole énoncée
expression manifestée
expression manifestée
Manifestation
Manifestation
Manifestation
Rue des boutiques obscures / Patrick Modiano.
- [Paris] : Gallimard, 1978. – 1 volume
(213 pages) ; 21 cm. - ISBN 978-2-07-028383-5
Rue des boutiques obscures / Patrick Modiano.
- [Paris] : Gallimard, 1982 . – 1 volume
(250 pages) ; 18 cm . – (Folio ; 1358). - ISBN
978-2-07-037358-1
Rue des boutiques obscures / Patrick Modiano, auteur ;
Jean-Louis Trintignant, voix . - [Paris] : Auvidis, P 1987.
- 4 cassettes audio (5 h 15 min) : Dolby.
- Auvidis Z115 (Coffret) : AD 803

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.

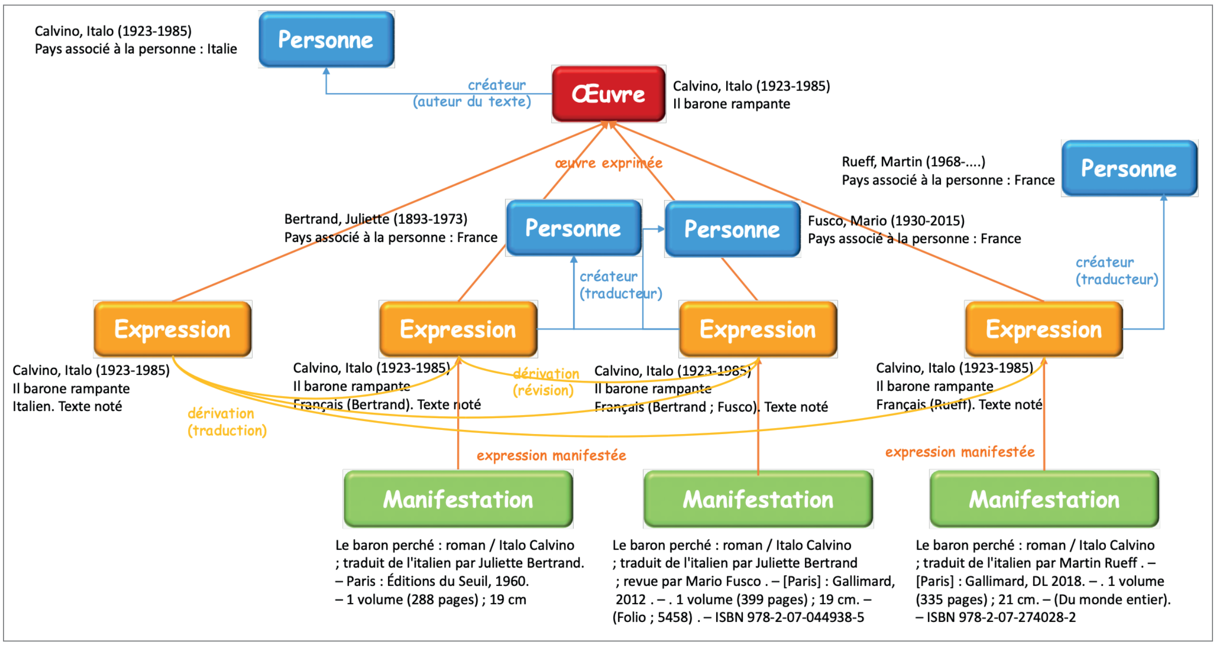

Calvino, Italo (1923-1985)
Pays associé à la personne : Italie
Personne
créateur
(auteur du texte)
Œuvre
Calvino, Italo (1923-1985)
Il barone rampante
œuvre exprimée
Rueff, Martin (1968-....)
Pays associé à la personne : France
Personne
Bertrand, Juliette (1893-1973)
Pays associé à la personne : France
Personne
Personne
Fusco, Mario (1930-2015)
Pays associé à la personne : France
créateur
(traducteur)
créateur
(traducteur)
Expression
Expression
Expression
Expression
Calvino, Italo (1923-1985)
Il barone rampante
Italien. Texte noté
Calvino, Italo (1923-1985)
Il barone rampante
Français (Bertrand). Texte noté
dérivation
(révision)
Calvino, Italo (1923-1985)
Il barone rampante
Français (Bertrand ; Fusco). Texte noté
Calvino, Italo (1923-1985)
Il barone rampante
Français (Rueff). Texte noté
dérivation
(traduction)
expression manifestée
expression manifestée
Manifestation
Manifestation
Manifestation
Le baron perché : roman / Italo Calvino
; traduit de l'italien par Juliette Bertrand.
– Paris : Éditions du Seuil, 1960.
– 1 volume (288 pages) ; 19 cm
Le baron perché : roman / Italo Calvino
; traduit de l'italien par Juliette Bertrand
; revue par Mario Fusco . – [Paris] : Gallimard,
2012 . – . 1 volume (399 pages) ; 19 cm. –
(Folio ; 5458) . – ISBN 978-2-07-044938-5
Le baron perché : roman / Italo Calvino
; traduit de l'italien par Martin Rueff . –
[Paris] : Gallimard, DL 2018. – . 1 volume
(335 pages) ; 21 cm. – (Du monde entier).
– ISBN 978-2-07-274028-2

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)

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.

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 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).

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

Remerciements à Emmanuel Jaslier (Bibliothèque nationale de France, Département des métadonnées) pour les informations sur la définition d'une politique de qualité des données à la BnF.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12766

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

Contact: Michele Casalini, michele@casalini.it

Received: 15 June 2021; **Accepted:** 21 June 2021; **First Published:** 15 January 2022

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12766

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

(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

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.

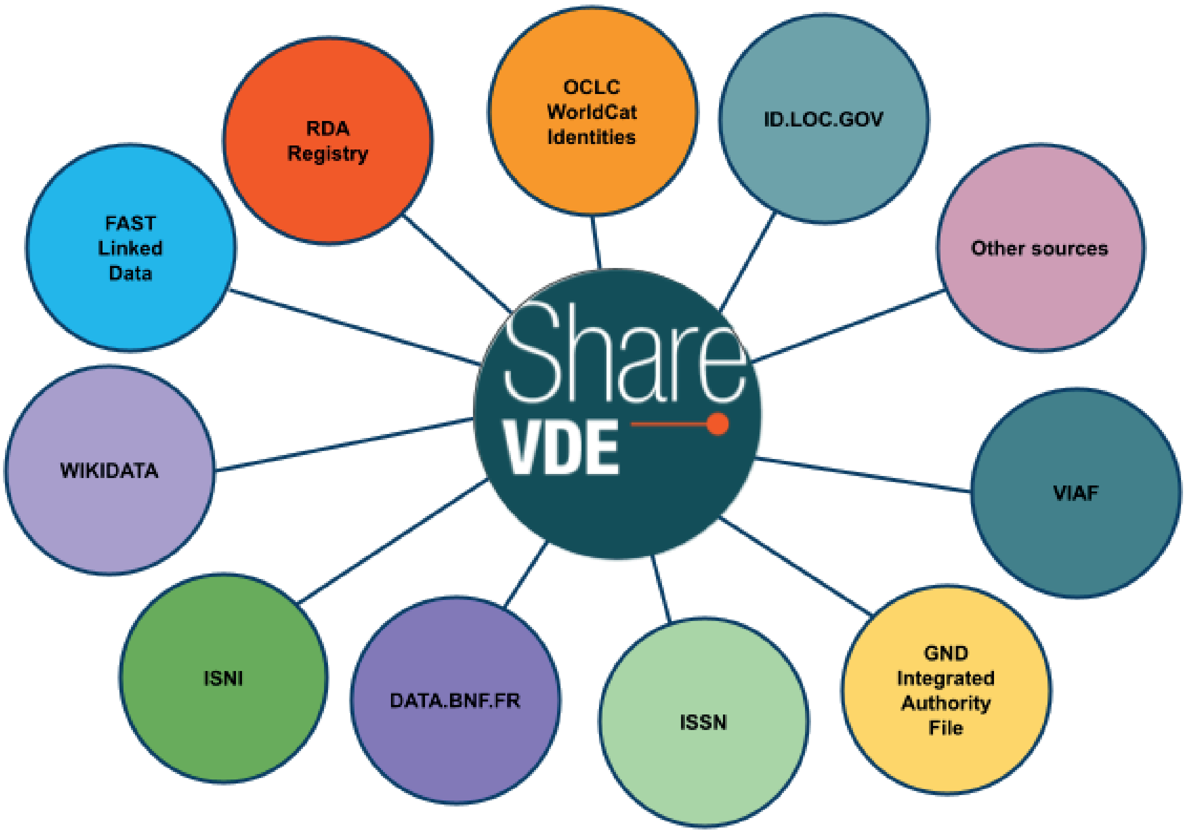


Fig. 1. Integration of data from external sources

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.

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.

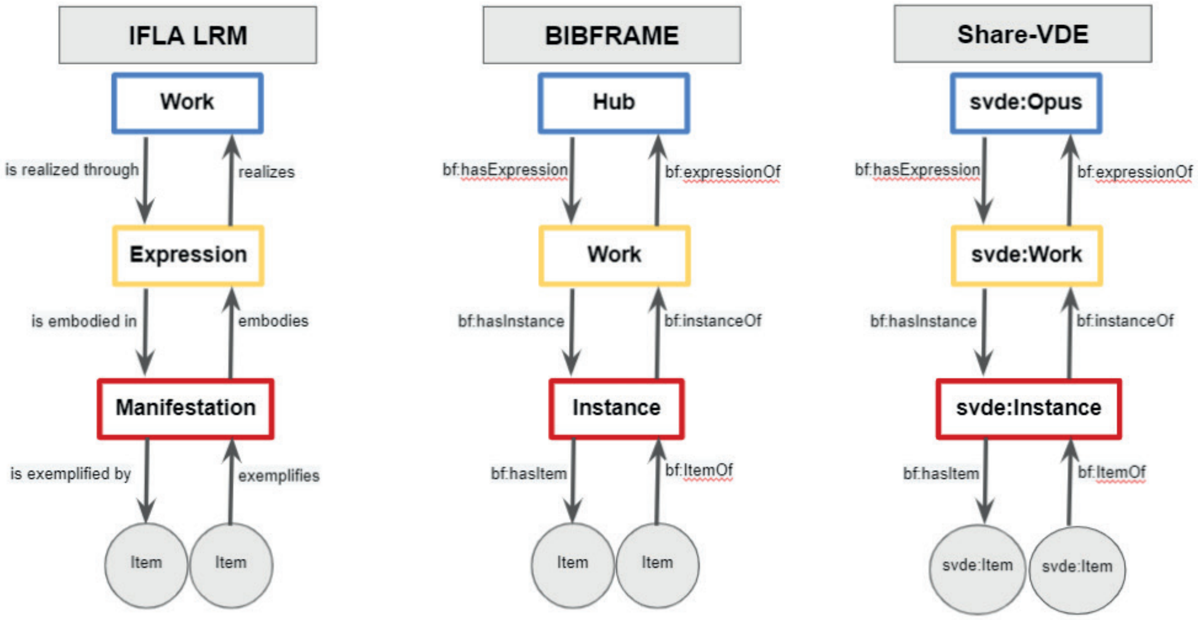


Fig. 2. IFLA LRM, BIBFRAME, Share-VDE Data model comparison. Further details on this structure can be found on https://wiki.svde.org

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.

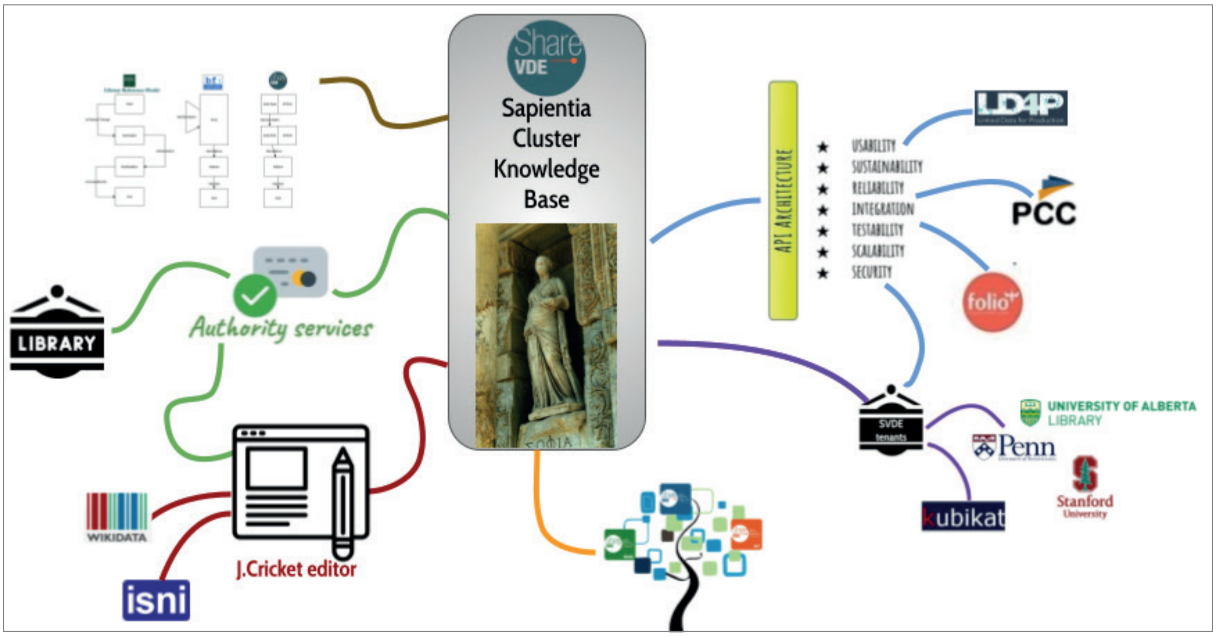


Fig. 3. Sapientia Cluster Knowledge Base ecosystem

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 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.

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

This contribution aims to offer insight into some of the practical aspects that have emerged from the experience of elaborating, experimenting and bringing new forms of authority control into operation.

It is also based on the cooperation and active contributions of many colleagues both in the library world and within Casalini Libri and @Cult, to whom I am extremely grateful.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12777

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

Contact: Piero Attanasio, piero.attanasio@aie.it

Received: 25 August 2021; **Accepted:** 12 September 2021; **First Published:** 15 January 2022

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12777

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).

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.

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.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12777

- 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.

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.

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12752

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

Contact: Fulvio Guatelli, fulvio.guatelli@unifi.it

Received: 15 May 2021; **Accepted:** 7 June 2021; **First Published:** 15 January 2022

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12752

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-

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12752

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.

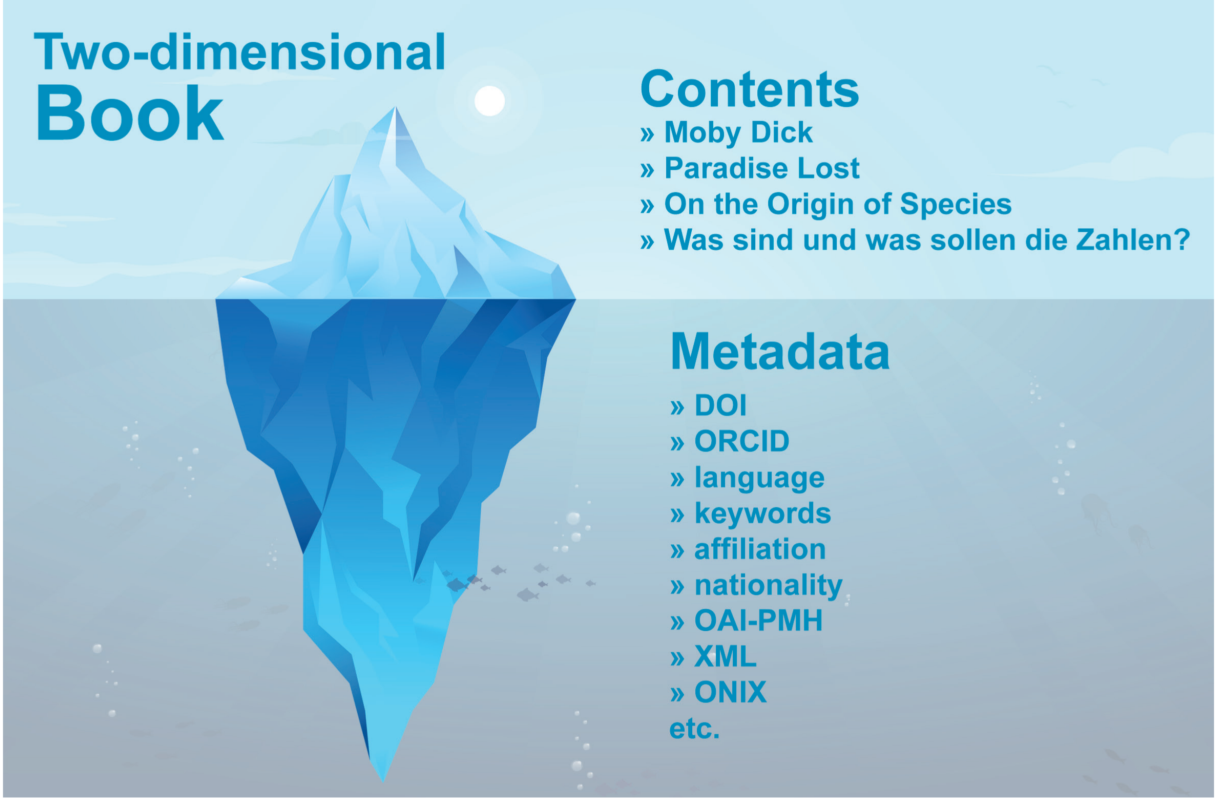


Fig. 1. Icebergs and two-dimensional books (CC BY 4.0)

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-

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.

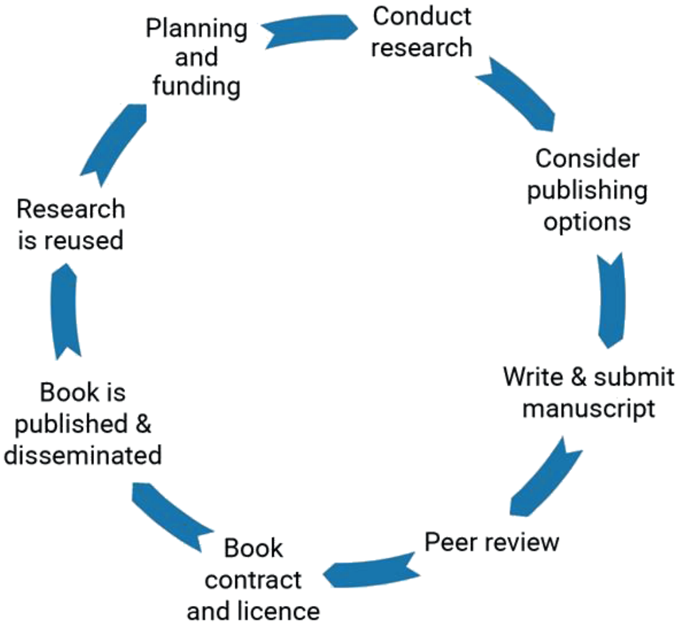

Planning and funding
Conduct research
Consider publishing options
Write & submit manuscript
Peer review
Book contract and licence
Book is published & disseminated
Research is reused

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 vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12752

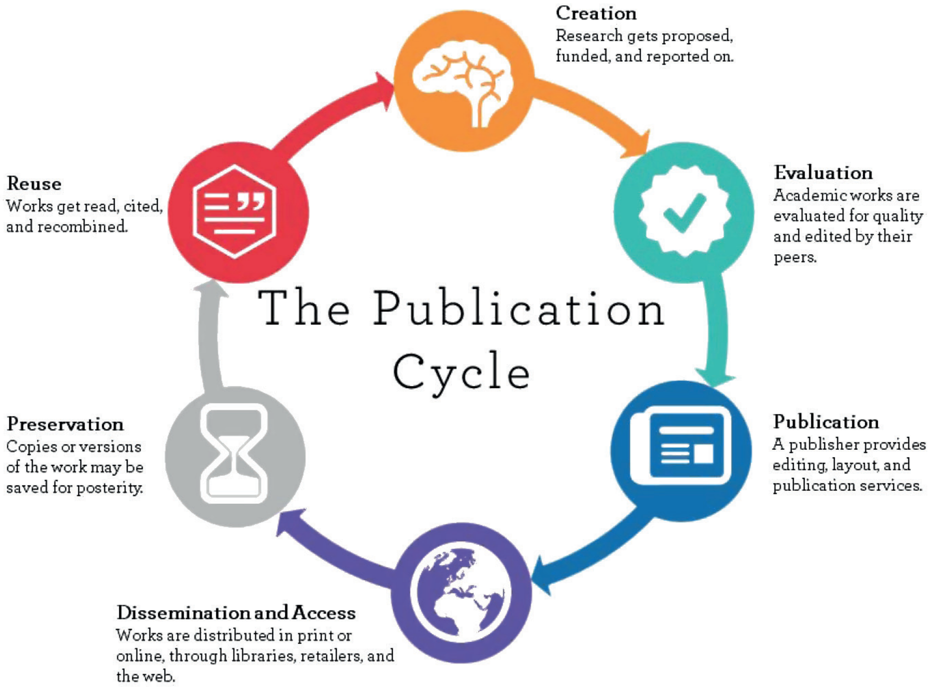

Creation
Research gets proposed, funded, and reported on.
Evaluation
Academic works are evaluated for quality and edited by their peers.
Publication
A publisher provides editing, layout, and publication services.
Dissemination and Access
Works are distributed in print or online, through libraries, retailers, and the web.
Preservation
Copies or versions of the work may be saved for posterity.
Reuse
Works get read, cited, and recombined.
The Publication Cycle

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:

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12752

- 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, OAIster, 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).

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.

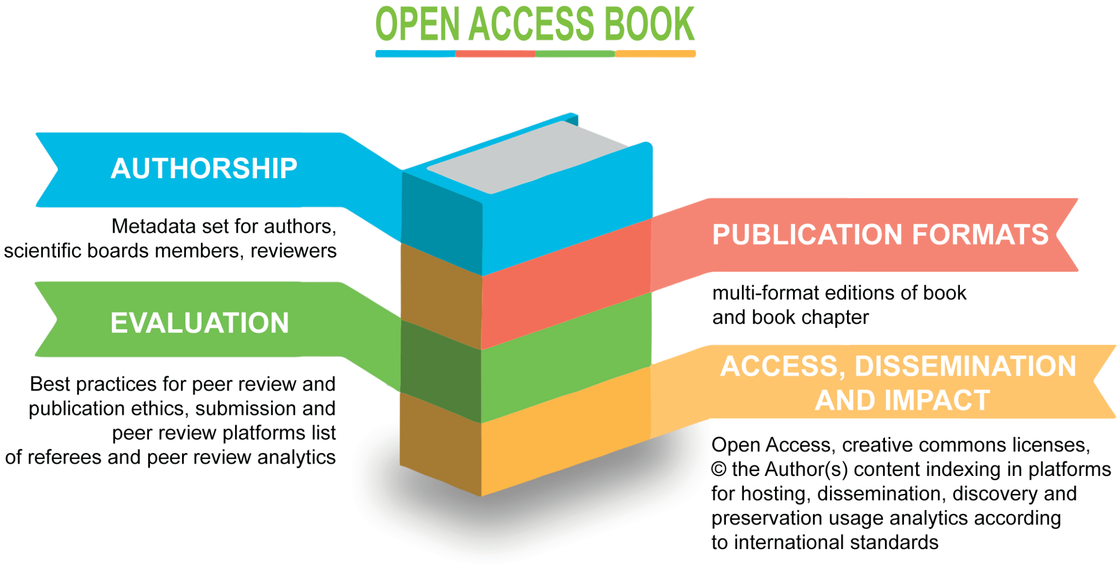

OPEN ACCESS BOOK
AUTHORSHIP
Metadata set for authors, scientific boards members, reviewers
EVALUATION
Best practices for peer review and publication ethics, submission and peer review platforms list of referees and peer review analytics
PUBLICATION FORMATS
multi-format editions of book and book chapter
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)

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.

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.