ISQ

TOPIC

EVOLUTION OF BIBLIOGRAPHIC DATA EXCHANGE

NISO
How the information world
CONNECTS

IP

Gildas
Illien

Interview with Gildas Illien,
Director, Bibliographic and Digital Information Department, Bibliothèque nationale de France (BnF)

Guest Content Editor, Ted Fons, presented Gildas Illien with a series of questions about the work that the Bibliothèque nationale de France is undertaking to transform bibliographic data exchange and to get his insight on the trends in the European library environment.

Q Can you summarize your opinion about the need for a new framework for bibliographic data exchange? Why is it necessary now? What is the biggest problem that we need to solve as metadata professionals?

There are many drivers for change in this area. Speaking from a national library perspective, I should start by stating that in a world where public policies and federal or national agencies are being constantly reevaluated, libraries need to demonstrate more output from their cataloging and metadata computing effort than they may have done in the past. The structured information they produce at high cost is expected to be used by more users from communities not restricted to libraries and linked to other data types in order to produce new knowledge and new services to people. The historical mission of national bibliographies remains valid in principle but must be radically revisited within this broader perspective of extended usage in the context of Linked Open Data.

I tend to look at the bibliographic transition we are undertaking mainly from a management angle—which is my role and contribution at the BnF [Bibliothèque nationale de France]. In this position, my first concern is to evaluate whether our cataloging workforce is successful in serving what should remain its ultimate purpose: access and usage. In this respect, it has become commonplace to acknowledge radical shifts in information research and retrieval practices. Our end users are on the web. They are looking for relevant and trusted information more often than they are looking for specific documents. Fewer and fewer search bibliographic information specifically, nor within the particular boundaries, languages, and applications of library catalogs. Things, People, Places, and Dates need to be expressed in more generic terms and concepts matching web standards and practices. Moreover, when it comes to researchers or corporate organizations, we know their need is no longer about finding and reading documents only, but also about confronting and mining large (meta) datasets using new computing tools.

THINGS　　　　　PEOPLE　　　　　PLACES　　　　　DATES

05.12.14

Things, People, Places, and Dates need to be expressed in more generic terms and concepts matching web standards and practices. Moreover, when it comes to researchers or corporate organizations, we know their need is no longer about finding and reading documents only, but also about confronting and mining large (meta) datasets using new computing tools.

Many librarians feel they are competing with or being defeated by the web, while they should see this new environment and expectations as a great opportunity to promote the information they've been producing and managing for decades: unlike much of what one finds online, bibliographic information is standardized and worthy of trust. National libraries and bibliographies possess an amazing legacy of highly structured metadata that could make a difference in making the web smarter. Library data may look complex from the inside (and actually is) but this complexity, if properly used, could improve the search and discovery end-user experience. In my view, our first priority should thus be to make bibliographic information fully interoperable with the web standards and environment, especially those of the Semantic Web. Bibliographic data exchange transition must be envisioned within this global and digital context, which should certainly have a strong impact on the data exchange modeling and infrastructure we will choose.

Economic constraints (budget and staff cuts) and the continuous growth of the amount and types of publications (both analog and digital) libraries are mandated to handle bring a second driver for change. To summarize what many experience these days, libraries need to do more with less. They can no longer afford the luxury of duplicating efforts and have to rely on much more cooperation, with a variety of stakeholders. While focusing their domestic production effort on their added value and unique or rare collections and references which will enrich the "long tail" of web contents, they will need to aggregate, confront, match, merge, or link an increasing amount of heterogeneous metadata from various provenances and of different status and quality level. As a result, institutions will have to organize many more data interactions and workflows involving other parties: interactions between libraries of course, but also between libraries and publishers along with other communities such as archives, museums, or research institutions. Many libraries may also consider giving a fresh eye at crowdsourcing in metadata, which will require managing direct interactions of end users with their bibliographic data or bridging their activities with those of powerful collaborative entities such as Wikipedia.

This means that from the original creation of records, metadata specialists will have to evolve as they will be handling more and more tasks designated to import, export, and transform metadata rather than creating it. This may imply outsourcing some of these tasks, sold as services by vendors, and participation in regional or global initiatives and knowledge bases maintained in the cloud for datasets that will not necessarily be made available for free. On the other hand, national libraries and bibliographic agencies will

need to remain worthy of trust and to maintain public, sustainable and free access to the databases they produce. Standards of quality, transparency, and publicity of the metadata they publish are crucial values they are certainly not ready to give up. This particular tension is to be taken into consideration as well in our vision of future metadata exchanges. There are and there will be even more players than today in the data arena, all with diverse, sometimes conflicting interests, missions, and business plans. In my view, the discussion of the business models capable of accommodating these various interests is also part of the picture we need to keep in mind while designing new data exchange infrastructure schemes.

I can see many technological opportunities to address these issues now, and to take action accordingly. The web of data quickly develops, offering potential solutions to some of these problems, provided professionals accept to move away from library-centric schemes and formats in order to seek better interoperability in a larger environment. It is *now* that libraries need to take position within the web of data if they want to be considered as significant players in this new environment—later might be too late. This is why we need to massively publish vocabularies and bibliographic data now, even if they aren't as perfect as we would like them to be. From a metadata specialist perspective, I would say the biggest problem underlying all these issues may be: how much are we ready to give up, as libraries, from our added value, from our legacy, from our specificities in order to accommodate such interoperability needs? I believe we will certainly need to change most of our cataloging habits, standards, and tools—which certainly are crucial attributes to a cataloger's culture and professional identity—but that losing the quality and granularity of the data itself should not be a requirement. What we need to do is to reformulate the information we manage in different terms. In the past 40 years, be it with MARC or other formats such as Dublin Core, we have experienced the limitations of trying to answer all functional and community requirements with a single format or implementation scheme. One size can't fit all and doesn't need to. The international community should rather consider developing strategies where various approaches may co-exist.

I would say we are ideally looking for a scenario where we could meet the joint requirements of:

a internal metadata management, including the management of legacy data not only for descriptive purposes, but also for digitization, rights management, and long term preservation of collections;

b rich bibliographic data exchange services with no loss of granularity in description; and

c standard data exchange and exposure on the web the people and search engines use.

Q What has the BnF already done to transform the way you express your bibliographic data?

I think our first challenge in the past years has been to change our general vision and strategy as to the bibliographic transition and to adopt a more pragmatic, perhaps more relaxed attitude as well, finding the right balance between international interoperability dependencies and the need to demonstrate tangible

> Data.bnf.fr does not mean to replace the existing catalogs and other silos it exploits, but to provide some "glue" between them. In short, it aims at making our library data work better on the web, by delivering a service of information, with structured explicit data and permanent URIs.

progress internally and at the national level. We felt the need for change but our initial vision to move forward was very linear. Initially, there was an assumption that, to do things properly, we first had to change the cataloging rules and standards, then envisage actual change of practices and tools for production. It was only at the end of this tunnel that we would eventually envisage how this long-term process impacting many people and involving considerable investments would practically make a difference to the end user. This was too stressful and too risky a process, also a very difficult roadmap to sell to our stakeholders and decision makers.

The BnF is investing heavily in the standardization effort and its best metadata specialists are still very much involved in ISBD, RDA,[1] and FRBR work, together with the national and international community. However, we are now looking at things the other way around. Our current priority is to work on the actual diffusion of our legacy data in order to achieve convincing and visible results in terms of web exposure and service. This has involved launching large data transformation campaigns of our catalogs, and supporting innovation efforts through various channels, always following the FRBR principles. Launching proofs of concepts, evaluating them, analyzing usage and community feedback, then scaling and industrializing them if relevant is currently our preferred method for organizing the transition. We learn and decide by doing and according to opportunities we discover step by step, while trying to take consistent options in the long run. When the benefits of change will become obvious to the majority, we will be able to change the production methods and infrastructure.

The main visible manifestation of this approach is the data.bnf.fr project.[2] This application was designed to be usable by individual, human-driven browsers, navigating through the various pages of a website. It generates web pages providing standardized information, references, and links about authors, works, or subjects. The service is also intended to be used by machines and search engines in particular. Data.bnf.fr groups and exposes online data in RDF form coming from heterogeneous sources which can be easily indexed by search engines and densely linked to other resources, either internal to the BnF (its MARC and EAD main catalogs, the digital library Gallica, etc.) or external (the

Union catalog for French Academic libraries SUDOC, the French Union catalog CCFR, WorldCat, VIAF, Wikipedia, etc.). The whole process requires the transformation of MARC or EAD formatted metadata into the information hub, based on modeling techniques in RDF and on standard vocabularies (DC, SKOS, RDA, and FOAF). The modeling activity has a direct link with aligning and enriching the data that have to be extracted and processed. Contents, links, and services are brought together in compliance with information concepts based on the FRBR bibliographical entities or groups of entities: those are integrated within a publication architecture designed both to build the HTML pages and to display raw data dumps in RDF and JSON. The data gathered from various datasets is brought together at the right level, so that works and expressions can be found in a way that complies with the new bibliographic description requirements. Data.bnf.fr does not mean to replace the existing catalogs and other silos it exploits, but to provide some "glue" between them. In short, it aims at making our library data work better on the web, by delivering a service of information, with structured explicit data and permanent URIs—a bibliographic information hub constitutive of a trusted environment made of reliable data. In order to facilitate data dissemination and reuse, all raw datasets are made available for free download under an ODC-BY and CC-BY compliant public open license recommended by the French Government Open Data mission Etalab.

Launched in 2011, this project demonstrates encouraging results. With over 5.6 million links to bibliographic records from the BnF main catalog, covering 200,000 authors, 92,500 works, and 171,000 subjects or themes, it is now estimated to cover 40% of the references from the BnF source catalogs. We target to reach 80 to 90% of the total by the end of 2015. At the end of 2012, for its first full year in operation, data. bnf.fr cumulated 637,650 unique visitors and 1.2 million page hits. On a monthly basis, we currently observe an average of 50,000 unique visitors per month. 80.6% of the visits come from a web search engine. This is an encouraging figure, which shows that most people using data.bnf.fr find it via a search engine, demonstrating success as to web exposure. The conversion rate is 70%, which means that 70% of the visits to data.bnf.fr lead to a visit of another BnF application (catalogs, Gallica, etc.). This is a good figure as well, as it

---

[1] To follow the discussions and ongoing work on RDA implementation in France (strategy, standardization, education, dissemination...), see this dedicated website (in French) "RDA en France": http://rda-en-france.enssib.fr/

[2] Data.bnf.fr won the 2013 Stanford Prize for innovation in research and national libraries. The text supporting BnF's application for this prize provides a comprehensive presentation in English of the project goals and outcomes: http://library.stanford.edu/projects/stanford-prize-innovation-research-libraries-spirl/2013-spirl-winners
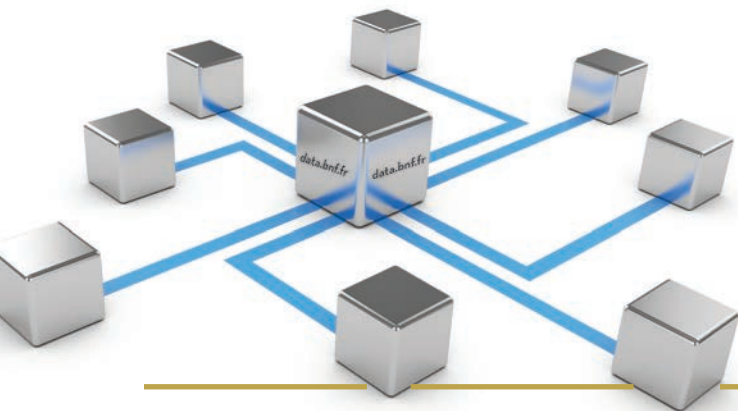
The conversion rate is 70%, which means that 70% of the visits to data.bnf.fr lead to a visit of another BnF application (catalogs, Gallica, etc.). This is a good figure as well, as it shows that data.bnf.fr is fully playing its role as an information hub (rather than a substitute), driving new traffic towards other BnF resources and applications.

shows that data.bnf.fr is fully playing its role as an information hub (rather than a substitute), driving new traffic towards other BnF resources and applications. The BnF cataloging staff has shown great interest in the development of this project. It is indeed a very concrete use case for professionals to see the data they produce in MARC presented in FRBR mode. The project development leads to the discussion of priorities and processes in the bibliographic transition: which datasets should be exposed next in data.bnf.fr and along which quality or content criteria? To which external data should the BnF link its own data to? Should current data transformation processes and algorithms influence the existing metadata models and production practices? Conducting such conversations and encouraging collective decision-making on the basis of this project has considerably improved the general perception of metadata issues at the library.

Although this project is used as a powerful vehicle for internal and external communications, it is only the visible part of the BnF bibliographic iceberg. Behind and beside data.bnf.fr and the continuation of our long-term effort in standardization work on ISBD, RDA, and FRBR, we have identified some "building blocks" that we believe will be key requirements to sustain the library's future presence and architecture in the Semantic Web. One of these building blocks has been the implementation of a comprehensive approach for the management of persistent identifiers. This started several years ago by assigning ARK identifiers to all objects and records from the library. Our current priority in this field is the implementation of ISNI for public identities. We are convinced that the management of authorities should be a strong focus to prepare the future, which explains BnF's strong institutional involvement both in the VIAF council

and in the ISNI Agency. Last summer, we managed to ingest 1.3 million ISNI identifiers in the BnF catalog. We are now planning their dissemination via our bibliographic services and are hoping this will ultimately answer some of the expectations of French publishers, academic, or rights management organizations, which are all in need of a global identifier to manage information databases about creators. In this context, we are getting closer than in the past to French publisher organizations, seeking more interoperability solutions and envisioning new workflows between their publishing industries and the library within the legal deposit framework, notably its extension to e-books. This involves working on ONIX/INTERMARC conversions and exploring various scenarios where the BnF could derive more metadata from the publishers just like we now derive many more records from WorldCat for our foreign acquisitions.

As to metadata exchange, our observation so far is that the data model designed for data.bnf.fr seems to be an acceptable compromise between generic web usage and exposure and basic bibliographic exchange needs: it is poorer than MARC but richer than schema.org, which we use in data.bnf.fr but consider more like a sitemap for webmasters and search engines than a data model. However, the way we serve our metadata in data.bnf.fr is not rich enough for high quality bibliographic exchange. This is why we are now looking into the possibility of expressing the full granularity of our INTERMARC format in RDF, the goal being to offer triple stores (via SPARQL endpoints) where people could just pick and choose what they need.

Q Can you summarize the focus of European libraries in the past five years? What has been the main focus of effort under the topic of metadata management?

The current discussions developing in North America and within the broader Anglo-American cataloging community regarding bibliographic data exchange models raise a mix of excitement and confusion in Europe. In the view of many European national libraries and bibliographic agencies, the invention and consolidation, within the framework of IFLA, of the FRBR model (and its later extensions to authority records and subjects with FRAD and FRSAD) is seen as the conceptual starting point of what we now call the bibliographic revolution. It is a strong view in Europe that the vision underlying this model remains valid and should be the main driver for bibliographic change, as FRBR is being consolidated by the IFLA international principles of cataloging while allowing for innovation and adjustments to the digital Age. European libraries invested a lot in FRBR theory and data modeling and still do, as shows, for instance, current developments with FRBRoo and PRESSoo and other models deriving from FRBR.

The development of the RDA cataloging rules and the beginning of their actual implementation in several major libraries is seen as a very significant and positive step to implement the FRBR model and make it happen in real life. Several European libraries, mainly from the AACR2 and MARC 21 tradition, have started translating and implementing it or are planning to do so in the coming years. Others, coming from different bibliographic traditions—mostly ISBD and UNIMARC—still see some limitations in RDA and aren't eager to adopt it as it is, mainly because it doesn't fit some of their practices and still requires some improvements in terms of internationalization or full compliancy with the FRBR model. From that perspective, considering the cost of change, there is a notion that if they should invest in such radical change, it should be for ambitious implementation scenarios which best fit the promise of FRBR. These institutions have put much effort in understanding the rationale of RDA and proposing adjustments where they needed them. The European RDA Interest Group (EURIG) was formed two years ago to provide a forum for European bibliographic organizations to collectively discuss and propose adaptations to the RDA code in order to address these issues. To date, this process and the subsequent interactions with the RDA Joint Steering Committee have been judged a constructive one, where all parties are given a voice. Although the whole process can be too slow, we know international standardization in the bibliographic field is one of the most complex types and that consensus cannot be achieved in one day in such matters. All in all, the dynamics of moving from the FRBR model to the RDA rules and their actual implementation following principles of international cooperation are regarded as a very encouraging process in Europe. Most European libraries seem ready to make compromises in order to reach some agreement so that institutional roadmaps may converge in the same directions for the benefit of international interoperability and future metadata exchanges. This is the exciting part.

The more confusing part has to do with recent developments regarding data exchange models in North America. Several European libraries perceive a contradiction between the collaborative effort which helped in designing FRBR and RDA over time and the way the question of data infrastructure is presently being addressed. While both FRBR and RDA are supposed to be agnostic as to

In the view of many European national libraries and bibliographic agencies, the invention and consolidation, within the framework of IFLA, of the FRBR model (and its later extensions to authority records and subjects with FRAD and FRSAD) is seen as the conceptual starting point of what we now call the bibliographic revolution.
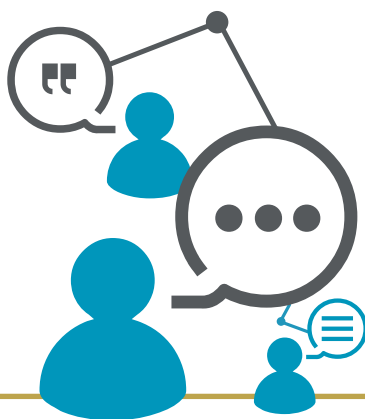
technical implementation, there is an overall feeling—which might be more of a misunderstanding about what the BIBFRAME initiative is actually trying to achieve—that important decisions and standards may be defined overseas without sufficient discussion with European libraries nor in compliancy with the initial vision and objectives that led to the definition of the FRBR model. At this very stage, I would say that this situation is a source of confusion for many, especially in the context where libraries feel the urge of demonstrating tangible results in metadata transformation and in developing new services fitting the Linked Open Data legal and technical requirements. Some libraries have started making their data open, but the data isn't linked. Others have started linking their data, but it's not open. Nobody really knows if the data exposed in RDF is being reused or has found proper metrics to evaluate this. FRBRization experimentations are being conducted in catalogs, at various levels of ambition, and through various channels (whether encouraged by ILS vendors or run internally via specific projects).

> I personally believe that we need to restore the conversation within the international bibliographic community and to encourage better communications between metadata standard specialists and linked data architects. This could help clarify things and avoid some misunderstandings.

Q **What should be the focus of the new metadata initiatives in the next two years? Are there any gaps in the current efforts that could be filled in the near term?**

All in all, there is currently a bit of confusion on how various institutional, national, regional, and global initiatives may converge as it seems to me that there is no proper framework to share best practices and confront technical implementation with standards requirements. This is all the more critical since within institutions it is often not the same teams who are involved in bibliographic standardization and in linked data projects, which makes it rather challenging to identify institutional policies or strategic roadmaps. This is an issue each institution should try to address internally.

In the meantime, it seems to me that while we had a rather clear focus and collective framework on the basis of FRBR and RDA in the past years, the urge of action has lead North American as well as European organizations either to act individually or to adopt a "wait and see" attitude which sometimes paralyzes them, especially when they are short of resources. I personally believe that we need to restore the conversation within the international bibliographic community and to encourage better communications between metadata standard specialists and linked data architects. This could help clarify things and avoid some misunderstandings. Typically, many people (especially at management level) tend to mix up models (e.g., FRBR), cataloging rules (e.g., RDA), formats and languages (e.g., MARC or RDF), and technical implementation solutions while these concepts operate at different levels, in different timeframes, and have different impacts.

It is obvious that different strategies will develop around the world as to the bibliographic transition, depending on institutional priorities, legacies, dependencies, and resources—especially in the context of the Semantic Web, which precisely allows for a diversity of approaches. But it could be helpful to define core areas of cooperation and implementation. Among those "building blocks" for the future that may require more international cooperation and reciprocal benchmarking (using existing forums such as IFLA or DCMI, or creating new,

dedicated platforms), I believe we should list: data exchange models, licensing and legal issues, publication and alignments of vocabularies, and global identifiers. At this stage of the bibliographic transition dynamics, we would benefit from a shared vision on these issues, which would help institutions planning their actions with a better notion of the areas where strong interoperability aspects are to be considered (and consensus searched, by means of collaborative discussions on standards) and other areas where they should feel more comfortable doing what they want to do depending on to their specific needs and mandates. | IP | doi: 10.3789/isqv25no4.2013.05

Bibliothèque nationale de France - ISNI 0000 0000 7409 4530

It is obvious that different strategies will develop around the world as to the bibliographic transition, depending on institutional priorities, legacies, dependencies, and resources—especially in the context of the Semantic Web, which precisely allows for a diversity of approaches.

Bibliothèque nationale de France
www.bnf.fr/en/tools/a.welcome_to_the_bnf.html

BnF Data Project
data.bnf.fr

Dublin Core Metadata Initiative (DCMI)
www.dublincore.org

Dublin Core Metadata Elément Set
www.niso.org/standards/z39-84-2005/

Etalab
www.etalab.gouv.fr/pages/Licence_ouverte_Open_licence-5899923.html

European RDA Interest Group (EURIG)
www.slainte.org.uk/eurig/

FOAF Vocabulary Specification
xmlns.com/foaf/spec/

FRBR Object-Oriented Definition and Mapping (FRBRoo)
www.cidoc-crm.org/docs/frbr_oo//frbr_docs/FRBRoo_V2.0_draft_2013May.pdf

Functional Requirements for Authority Data (FRAD)
www.ifla.org/publications/functional-requirements-for-authority-data

Functional Requirements for Bibliographic Records (FRBR)
www.ifla.org/publications/functional-requirements-for-bibliographic-records

Functional Requirements for Subject Authority Data (FRSAD)
www.ifla.org/node/5849

INTERMARC
www.bnf.fr/fr/professionnels/format_intermarc/s.intermarc_presentation.html

International Federation of Library Associations and Institutions (IFLA)
www.ifla.org/

International Standard Bibliographic Description (ISBD)
www.ifla.org/publications/international-standard-bibliographic-description

International Standard Name Identifier (ISNI) International Authority
www.isni.org

JavaScript Object Notation (JSON)
www.json.org

MARC 21
www.loc.gov/marc/

ONIX
www.editeur.org/8/ONIX/

PRESSoo (a periodicals object-oriented ontology)
www.issn.org/files/issn/technicals/PRESSoo_01.pdf

Resource Description and Access (RDA)
www.rda-jsc.org/rda.html

Resource Description Framework (RDF)
www.w3.org/standards/techs/rdf#w3c_all

Simple Knowledge Organization System (SKOS)
www.w3.org/2004/02/skos/

SPARQL
www.w3.org/standards/techs/sparql#w3c_all

UNIMARC
www.ifla.org/publications/unimarc-formats-and-related-documentation

Virtual International Authority File (VIAF)
viaf.org

WorldCat
www.worldcat.org/

RELEVANT
LINKS