Linked open data for new library services: the example of data.bnf.fr

Romain Wenz

JLIS.it

Linked open data for new library services: the example of data.bnf.fr

Romain Wenz

Library catalogues were designed to locate books and to handle collections. They are used by librarians collecting books and by users finding them. Yet, it can be hard for a user to reach library information on the web, especially as there can be several catalogues for one library. Indeed, different kinds of tools are required for different kinds of collections. For instance, a collection of archives and manuscripts needs a hierarchical structure, to describe documents together, as they were produced and received during the activities of a person. Therefore managing documents can be different goal from making access to them. Web users have new expectations and new habits in a changing web environment. Library data should meet these needs and truly belong to the web. Libraries try to make their data really useful on the web. We will focus on the use case of data.bnf.fr, a project from the Bibliothèque nationale de France that relies on efficient links, automatic techniques and semantic web tools.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5509

New expectations

Online catalogues make things different. With the world wide web, researchers have access to plenty of resources from a single computer, even from home. There is now some kind of competition between document providers, since it is a lot easier to switch from one to another. For instance, a copy of book will be less needed once it is digitized and available online. The theories of Walter Benjamin showed that content is losing its value once it is copied with industrial processes, which is very true in the digital world. On the other hand online access creates another kind of value, at least for cultural and educational resources, which are meant to be spread. The resources provided by libraries have to be easy to find, because they become part of a more general "web search". There are always more documents online: many specific websites provide information which can be compared to what can be published in books. Moreover, digital collections published by libraries become part of the web. For readers who are looking for resources on the Internet, texts from digitized books provide information, like other web pages. That is why digital collections, but also online references of physical books, have to be easy to find and accessible through automatic programs. The general public can find some documents without even knowing they exist. Typically, with the use of powerful search engines, users now commonly search with keyword associated with the final document. This habit was spread in a decade. It has made all kinds of online information always easier to find, through the use of search engines. This implies that users tend to search with keywords associated together, as opposed, for instance, as using a series of fields as in catalogues. It also means that results sorted by algorithms are commonly accepted. We are all familiar with sentences such as "results 1 to 10 on 120000": noise is not a problem, if it comes after the relevant results, found automatically

and presented first. How should libraries take advantage of these evolutions? Several sources of information can help us, in order to decide how to adapt. First, the statistics of our local search interfaces provide accurate and free feedback on what our traditional users search. For instance, some years ago people used to search for "complete works" of writers, knowing in advance what they would find in the book. Now, we mainly have searches for the books themselves, typically with the title of the book and name of the writer. Public surveys from by the libraries or other institutions show that, using search engines for browsing the web has now become a habit. Internet users usually find bibliographic references online before going to the physical library. This is confirmed by all the user surveys made those past few years, for students as well as for researchers, as the ones made by OCLC,[1] and by the Bibliothèque nationale de France.[2] Therefore, book references that are impossible to find online are almost useless for most people. If librarians want them to be found, they have to put those references on the web. Most catalogues are available online with a specific portal. But they are usually not accessible from web link, and impossible to crawl for search engines. Those new expectations from the public are essential for libraries, because of the size of the content owned by libraries. The amount of content and information available is so huge that the most recent techniques have to be used to handle them. For instance, the Bibliothèque nationale de France displays 1,5 million objects in Gallica,[3] which is the biggest French-speaking digital Library, and 12 million bibliographic records, thanks to the legal deposit of the French edition. Thousands of manuscripts and archives are also available, with all types of resources, from medieval manuscripts to archival fonds of modern writers. Handling

[1] For instance http://www.oclc.org/reports/onlinecatalogs.
[2] For instance http://www.bnf.fr/documents/enquete_gallica_2011_rapport.pdf.
[3] http://gallica.bnf.fr.

this kind of resources creates several scale issues, as we are dealing with millions of documents. There are always duplicates, and the quality of the data is irregular, as a result of the long history of our catalogue. Moreover, printed books and manuscripts are usually described with various logics, inside the catalogues. Records from the main catalogue describe a physical book, usually in a MARC format. They are structured deliberately around a collection which was constructed on purpose, with a series of books that would be shelved together and make sense for the end-user. On the other hand, archival were produced and received during the activities of a person, and considered in a way as "by-products" of the life and activities of some person or organisation. The documents were gathered according to this logic, which is not always obvious for the end-user today. Therefore, the documents cannot be described with simple "records", but with the model of a hierarchical tree, which makes it possible to understand the original logic of the archives. The format which is commonly used for this kind of resources at the Bibliothèque nationale de France is XML-EAD (Encoding Archive Description). The digital collections, available in Gallica, are described with a simple format: Dublin Core. All digital items are accessible with a persistent identifier (ARK), given and maintained by the Bibliothèque nationale de France. Between these catalogues, efficient links have to be provided, so that the users can browse quickly and go simply from one document to another. Machines are not intelligent, so it is necessary to provide structured information in the catalogues, with efficient links between the documents.

Importance of efficient Links : principles in data.bnf.fr

If we want the resources to be truly part of the web, in the sense that users can quote them on sites, blogs, pages, and e-mails for instance, and access them by following links, we have to give them proper identifiers, and to comply with web standards. Thus it is also possible to link resources from our different datasets. As big libraries often have several catalogues, making links between them makes it possible to find resources without having to learn how to use all the different tools, just by "following one's nose". It makes it possible to handle library data at large scale, with different types of documents. This is very important since many distinctions between documents were made before the web. For instance, for the end-user, a digitized "regular book" and a digitized medieval manuscript can be equal, in the sense that the same user can access them in the same way if they are online. The very notion of "special collections" can change if they are digitized and available on the web. This form of openness is accomplished through digitization processes. In the context of digitization, many resources which were interesting only for specialized scholars have become relevant for a broader public. For instance, medieval miniatures are surprisingly used by a very broad public once they are online. The way to search has to be simple for these resources, open to the web with digitization. They have to be easy to find inside databases but also available through links on the world wide web. In general catalogues as well as for digital collections metadata, the data describing documents has to be technically available, but also legally re-usable if we want it to be broadly spread. This is why many libraries move to the techniques of the semantic Web, together with licenses of "open data". This way, some libraries are part of the "linked open data" movement

and are involved in the development of the "semantic web". The Bibliothèque nationale de France develops a new project, bringing together data from catalogues (MARC), archives (EAD) and digital resources (DC). All the data are extracted and gathered automatically. This project, called data.bnf.fr,[4] is still a young project, as it has been online for a year. Data.bnf.fr gathers descriptive information, and links directly to online catalogues and digital documents. There are several aspects: a first goal is to make the information compliant with the "semantic web" standards, by providing persistent identifiers for the resources, with a RDF view on the available information. For the library, gathering information around concepts of works, writers and subjects also implies to work on modelling issues. In fact, it is a first opportunity to implement the FRBR model, and to use it with automated matching and alignments. To do so, we use a free software, called CubicWeb.[5] This is not only a technical issue, but also a way to get a first feedback on what is possible, and how users react. Therefore, it is very important to publish both structured data for computers, and web pages, quick and easy to use for humans.

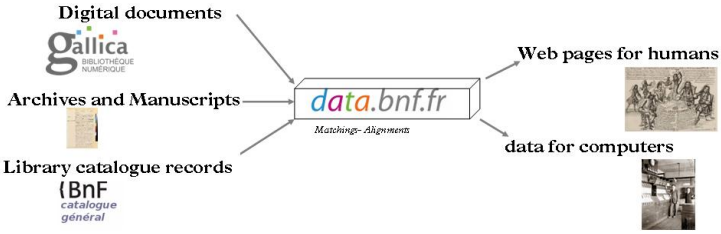

Figure 1: data.bnf.fr.

[4] http://data.bnf.fr.

[5] Site and documentation at http://www.cubicweb.org and http://docs.cubicweb.org.

So as to comply with the "semantic web" principles, it is essential to provide information which is described with common vocabularies, with a strict structure, the way they would be inside a database. This can only be done with identifiers given to all the concepts that have to be handled. The Bibliothèque nationale de France uses persistent identifiers for web URIs: the "ARK"[6] identifiers, which are used to identify catalogue records, archival resources, digital objects from Gallica, and authority records. These ARK identifiers are also useful for quoting these resources, with a common "resolver": for instance, a digital object[7] will also be accessible with a persistent link.[8] In order to gather information about Works, Writers and Subjects, data.bnf.fr relies on the authority files. The ARK identifiers that have been given to the authority records are also used for the pages in data.bnf.fr, so as to build reliable URIs, and efficient links between the data.

Use cases for Linked Open Data (LOD)

Concepts from authority files are basically used for reliable identification and description, for writers, books, and subjects. Then, the difficulty is to provide relevant links to editions of the books, manuscripts, archives, images. Some online examples can explain how data.bnf.fr copes with these issues. For instance, when searching for information about a writer like Goldoni,[9] the user will find all his main works with pages gathering all the editions of the books. There are links to the references of the editions of the catalogues, to the digital items, and to the manuscripts such as the letters which

[6] Archival Resource Key.
[7] http://gallica.bnf.fr/ark:/12148/bpt6k134521m.
[8] http://ark.bnf.fr/ark:/12148/bpt6k134521m.
[9] http://data.bnf.fr/11905320/carlo_goldoni.

Goldoni received. The authority files have been extracted, and the identifiers are used to make persistent links, and to avoid duplicates. The "Work" pages are created at the FRBR level. This means that the book is described as a concept and not as a particular edition of this Work. All the editions are gathered around this concept. For instance, on the page about the *Trionfi* by Petrarcha,[10] it is possible to find a list of the manuscripts and of the printed books, with links to all the digital items when they are available. Writers are of course linked with their works, with the associated documents, and with other writers. This kind of information is extracted from the documents linked to both of them. For a writer such as Leonardo Bruni, a page[11] gathers links to all editions, manuscripts, and digital items. The texts he wrote, translated, edited or commented are available separately, depending on his role on the document. The user can look for the translations he made, for instance. There are also links to the pages of the associated writers, such as Cicero and Aristotle. Since he was an editor of Cicero, we provide both the references of the edited books, and a link to the page of Cicero.[12] The web semantic tools and the reliable links enable us to build pages around the common properties and infer new relation from our RDF graph. For instance one can find:

- data.bnf.fr pages associated with the date 1515;[13]
- data.bnf.fr pages associated with the date 1789;[14]
- or all the authors who have been making coins, such as Louis XIV.[15]

[10] http://data.bnf.fr/11953648/petrarque_les_triomphes.
[11] http://data.bnf.fr/12027636/bruni_leonardo.
[12] http://data.bnf.fr/11885977/ciceron.
[13] http://data.bnf.fr/what-happened/date-1515.
[14] http://data.bnf.fr/what-happened/date-1789.
[15] http://data.bnf.fr/vocabulary/roles/r370.

Data.bnf.fr makes links and publishes web pages containing already about 2.5 million linked resources. The complete data is also displayed in RDF and available by clicking on the RDF icon, at the bottom of the pages, by adding the following suffixes to the URL: NT, N3, RDF-XML, according to the format needed,[16] via content negotiation, using a RDF web browser, from the URL, or by bulk downloads.[17] As yet (in the summer 2012), the complete available dataset is 6.3 million RDF triples, which is not to massive considering the 2.5 million resources, thanks to the proper links that avoid us too much redundancy. All the raw data is also displayed in RDF and available with an open license. Allowing all kind of uses, also for commercial purposes, was not obvious.

Why we use an open license for data.bnf.fr. Legal and technical requirements

Displaying information on the web means that the institution is responsible for publishing documents. Legally speaking, the library becomes responsible for the content which is displayed. From a "marketing" perspective, publishing information on the web is an incentive to focus on what you can do best, and to let others take care of the things they do better than you, because the users will prefer to use their resources anyway. For instance, library catalogues are describing resources, and handling "concepts" that have to do with documents. The strengths and weaknesses are not the same, for instance, as in an encyclopaedia. It would be no use to try to insert universal encyclopaedic knowledge inside library catalogues, just as it would be useless to provide full lists of documents in-

[16] Example http://data.bnf.fr/11928016/jules_verne/rdf.xml.
[17] From http://data.bnf.fr/semanticweb-en.

side general information encyclopaedias. On the web, libraries are bringing a long-time perspective. They have been collecting books for centuries and data for decades. The manpower provided has no equal in terms of describing books. Besides, the data has been structured quite early, with international standards since the 1960s. This "descriptive data" was not produced in a "marketing" perspective: all elements are accurate, and meant to be interoperable, even though several formats actually exist. The rules for producing the bibliographic descriptions remained stable, and were strictly followed by trained cataloguers. Through several formats for various types of documents such as books (MARC formats), archives (EAD), and digital resources (DC), the standards were respected. Therefore, the information can be trusted and processed automatically, even on a long period of time, and through huge amounts of data. The library catalogues are already machine-readable, even if it is not yet necessarily with web standards. Displaying them on the web with web standards implies to use identifiers (URIs) so that people can quote the resources. If we want to let people use these web resources, we must provide reliable links. Because this is a great opportunity to share our cultural materials, the Bibliothèque nationale de France decided to make the structured data from data.bnf.fr freely available, with an Open Licence. Opening library data on the web is a way to take part in the "open data" movement, and to give access to the information to the broader public, by using the most recent technologies. It is also an incentive for others to use this material and to give access to culture. By making the RDF data free, this project also take part of the international experimentations of "Linked Open Data" (LOD) that have popped up among national libraries. As Gildas Illien puts it, «Transforming pre-existing MARC records and authority vocabularies in RDF triples; starting to implement the FRBR model ; playing with the semantic web standards ; building

applications and datasets of a new, linked data-friendly type: this is what looking at LOD means to them at this stage» ("Are you ready to dive in? A case for open data in national libraries"). Because libraries are working on a long-term perspective, data.bnf.fr also tries to experiment on solutions that can be used in the original library catalogues. First, when developing "matching" and algorithms for gathering data around "Works", we try to provide honest information for displaying data on the web, for instance, to avoid having duplicates, to avoid displaying keywords that would not match with the content of the documents, or any other information that would in fact not be useful for the end-user. This is why we keep so many links to all the resources in the original catalogues, inside the pages of data.bnf.fr. We also try to build routines and mechanisms that can be used inside the original catalogues, in the long-term, for instance for automatically generating "Work" pages inside our authority files, according to FRBR. Besides, after a first year of presence online, we can already have a feedback from some users, on the kind of content that is being used. Some of them are re-distributing the dataset and referencing it for others to re-use, starting with data.gouv.fr,[18], the official open data portal of the French State, but also other sites such as CKAN[19] OKF[20] and open data directory.[21] Other users are data specialists from the cultural sector, who use a part of the data for specific purposes in their local applications, such as the *Institut français*.[22]

Some are developers who want to build timelines for research pur-

[18] http://www.data.gouv.fr/donnees/view/Donn%C3%A9es-compl%C3%A8tes-du-contenu-de-la-BNF-30383137.

[19] http://thedatahub.org/dataset/data-bnf-fr.

[20] http://en.wikibooks.org/wiki/Open_Metadata_Handbook/Technical_Overview#Biblioth.C3.A8que_Nationale_de_France_.28BnF.29.

[21] http://open.mflask.com/dataset/data-bnf-fr-bibliotheque-nationale-de-france.

[22] http://ifverso.com.

poses, such as "Yokafun",[23] or for Smartphone applications.[24] This broad range of uses of the "raw data" shows us that library information can be useful for broader communities, even if the first purpose remains to describe collections and to give access to them. When gathering all available resources at the level of intellectual "works", the dataset is not a "catalogue" in the traditional sense, because it is not necessarily used for identifying a document or handling a collection. It becomes part of the web, in a new way. The authority files and identifiers are more important than ever to build this kind of service, but the dataset itself is something else than a traditional catalogue. Besides, the web tools allow us to keep trace of the behaviour of users. We can of course collect the direct feedback on how people have been reacting to it, what kind of content is being used and what has to be improved, which leaves a wide range of possible improvements for the future.

Works cited

Illien, Gildas. "Are you ready to dive in? A case for open data in national libraries". *Libraries now! Inspiring, surprising, empowering. IFLA World Library and Information Congress. 78th IFLA General Conference and Assembly*. 2012. http://conference.ifla.org/sites/default/files/files/papers/wlic2012/181-illien-en.pdf. (Cit. on p. 413).

[23] http://plindenbaum.blogspot.fr/2011/07/drawing-svg-timeline-with-httpdatabnffr.html;https://gist.github.com/1093853.

[24] For instance https://sites.google.com/site/catbnf;http://www.appforcash.com/section/item/id/41491.

ROMAIN WENZ, Bibliothèque nationale de France.
romain.wenz@bnf.fr

Wenz, R. "Linked open data for new library services: the example of data.bnf.fr". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5509. DOI: 10.4403/jlis.it-5509. Web.

ABSTRACT: The Bibliothèque nationale de France (BnF) develops a new project, bringing together data from catalogues (MARC), archives (EAD) and digital resources (DC). It makes links and publishes web pages, available at http://data.bnf.fr, with already about 750.000 linked resources. All the raw data is also displayed in RDF and available with an open licence. The presentation will explain the importance of authority files and identifiers to build this kind of service, an give a first feedback on how users have been reacting to it: what kind of content is being used.

KEYWORDS: data.bnf.fr; Gallica; Bibliothèque nationale de France; Library linked data; Authority file

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15