Data.bnf.fr: an overall presentation

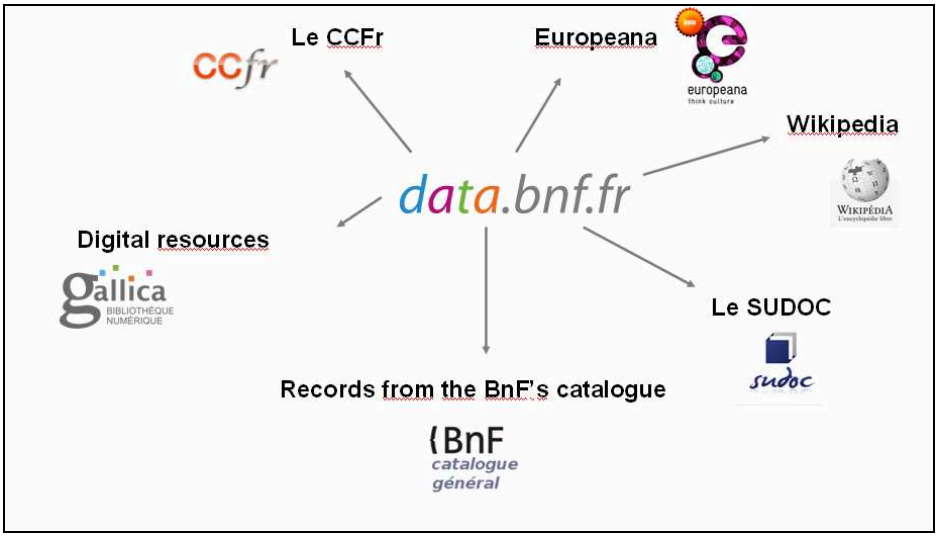
The Bibliothèque nationale de France has designed a new project in order to make its data more useful on the Web. It involves transforming existing data, enriching and interlinking the dataset with internal and external resources, and publishing HTML pages for browsing by users and search engines. The raw data is also available in RDF following the principles of linked open data architecture.

Keywords: Linked Data; Semantic Web; metadata; interoperability; RDF; URI;

1. Bibliographic data on the Web:

→ Putting forward BnF data

Library data can be difficult to find on the Web. At the BnF, it is of course possible to access all of the resources and services through our Library Website (www.bnf.fr). But, at present, few of them are indexed by search engines. And, even when they are, it is difficult to sort results from them.
Some digital books, even when they are completely and freely available, are sometimes impossible to find if you don't already know they exist. The data.bnf.fr project can be a way to open the digital library Gallica to a wider public. Moreover, library catalogues are usually stored as relational databases: they are just no use for Web search engines. Users always access the BnF catalogues (mainly, the Main catalogue and the Archive and manuscript catalogue) through library portals, which they often simply don't know. As a matter of fact, users are very unlikely to find any of our resources directly from a search engine interface, unless they already know about us.



Some links from data.bnf.fr.

Data.bnf.fr is a Web interface which gathers full digital document and descriptive data from different catalogues and enables users finding the relevant information in our resources.
Our resources should be as visible on the web as the library building in the town.

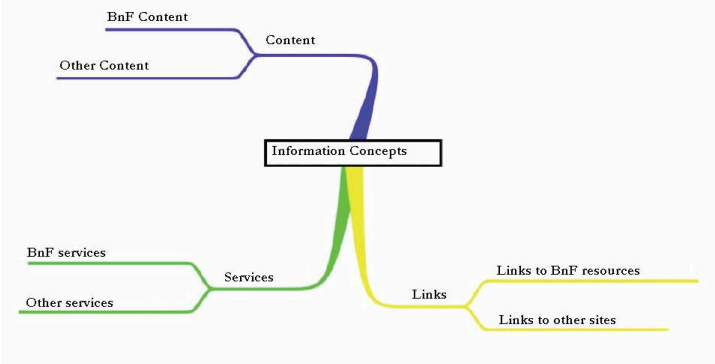
→ Structured Data have a value

Typed, normalized and labelled data is the basis of Web search. With record identifiers and labels, libraries already identify resources in a uniform way and "link data" Through the links between works, authors and subjects, librarians have been "linking data" for years. They have been providing useful and reliable information through authority files. Indeed, our library catalogue holds **more than 12 million records**, **all structured and linked together**. **It relies on two million accurate and trustworthy authority records about authors, corporate bodies, works and subjects (RAMEAU[1]), which are maintained, with permanent URIs (ARK identifiers[2] at the BnF[3]).** On the Web, data that are provided by a public institution such as a national library have a specific value,

[1] http://rameau.bnf.fr
[2] Bermes, Emmanuelle. (2006). *Les identifiants pérennes à la BnF*. Retrievable from http://bibnum.bnf.fr/identifiants/identifiants-200605.pdf

since they have no other purpose than to provide useful information, reliable sources, and quotable links. These ARK identifiers enable us to identify, quote, but also, to gather access to resources. Thus, we are able to align our resources, referring to the FRBR model (see below), always in order to bring new services to help users.

We want to provide the machines with the means to index access **to content, links and services,** for each page (Documentary unit) around a concept with a large meaning (see below). “Content” means descriptive, accurate and valid data, elaborated by a non-profit service. “Links” means a way to navigate and move to more relevant resources if necessary, particularly towards the online version of a work, and integration into a resource graph. “Services” can mean other library services, such as “ask a librarian”, download or print.



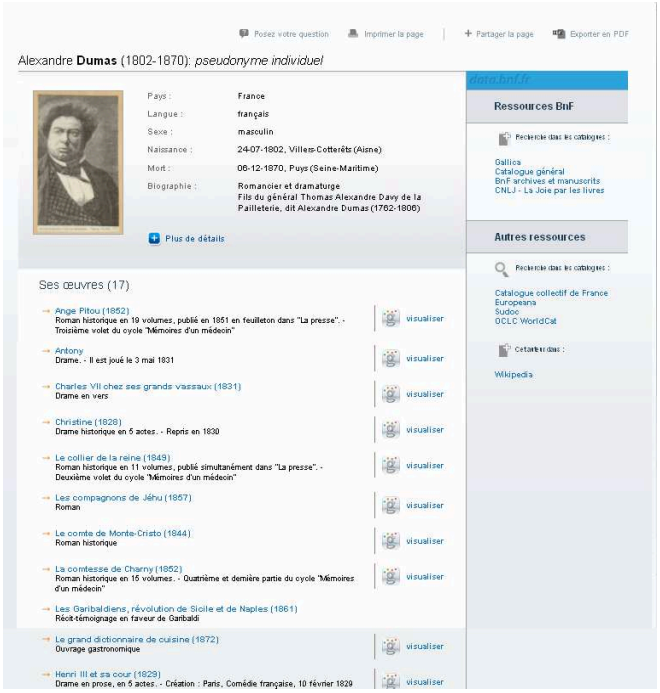
The website is structured around information concepts

Content, links and services are brought together in each page around an information concept. Data.bnf.fr is also based on modelling techniques and Resource Description Framework.

2. Web pages about authors, works and subjects

We have built a Web interface with html pages, gathering resources around the concepts of "works" and "authors". They are meant for a wide public. At the same time, we publish raw data with a model built around “concepts” and with interoperable data, which are exposed on the web of data. The basic issue for us is:

- on the one hand, how do we make sure we can answer frequent “short-term requirements” such as specific requests or strategic issues : resources that are popular at a time, or about graphic issues or fashionable tools, for instance.
- on the other hand, how to take into account traditional and long-term missions of the library, at the same time, such as providing technically advanced data models and solutions and valid and proper information.



Alexandre Dumas in data.bnf.fr

→ The link to the FRBR (Functional requirements for bibliographic records) model

This way to articulate bibliographic data on the Web implies several choices. As a matter of fact, the aim of publishing HTML pages implies that our data model will basically enhance concepts that are relevant for creating a Web page. We chose to rely on the concepts of works, authors and subjects, which happen to be entities in the **FRBR model**[4], **as we try to make our data model compliant with the FRBR requirements.** This Web interface is at the crossroads between the different resources we make available on the Web.
It gathers different kinds of data at the right level: works, expressions and manifestations. For an author, users find all the links to the Web pages of the relevant works, by and about the author, in two different sections. For a work, there is a link to the author's page, but also to the different manifestations of the work (bibliographic resources, online material). In order to create these pages, we need to bring data together from different BnF datasets, which are in various formats: EAD[5] (Encoded Archival Description) for manuscripts and archival fonds, MARC (Intermarc) for the main catalogue, Dublin Core[6] for the digitized book from Gallica[7] and for the virtual exhibitions[8]. Therefore the modelling activity has a direct link with aligning and enriching the data that have to be extracted and processed.

Finally, these pages provide a range of advanced functionalities (PDF export, export and send, quote on social networks…). Besides, there are links to other online services where the user is likely to find relevant information if the current page did not provide him with enough information. We retrieve data from other "open datasets" (such as Wikipedia) to improve matching and to provide another kind of information.

Data.bnf.fr's pages should:
- be easy to browse, search and find for the user
- develop or propose new models, such as the FRBR model or alignments with external datasets

3. Web of data:

We have built a publication architecture that enables us to have html pages and to display the "raw data" on the "Web of data" at the same time.

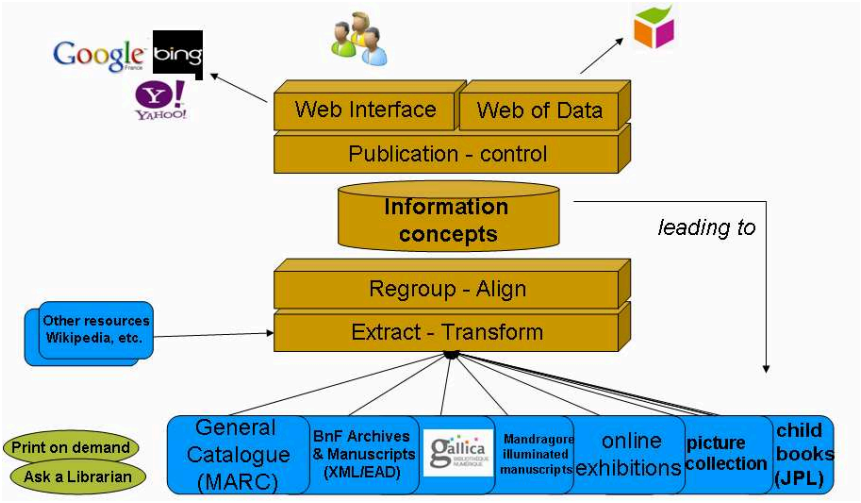
Our purpose is to use common standards, and to build this service through a "semantic-web" friendly data model which enables us to bring our resources and records into the "linked data", so as to make them as useful as possible for both library users and professionals.
By respecting the semantic web standards, we can bring structured data that are understandable and usable by machines and based on **interoperability not only with external sources, but also between our own different datasets, since we have to align resources from several catalogues.**
We also display the subject records (RAMEAU) from the French national library. They have been converted into the **RDF vocabulary SKOS** (Simple knowledge organisation system), in the context of the European project Tel plus. This repository has been **updated and completed with the current records from the BnF database**.
We use a software called CubicWeb, a semantic web application framework, licensed under the LGPL.

→ The requirements of the semantic web[9]

For the pages describing resources, we want to:

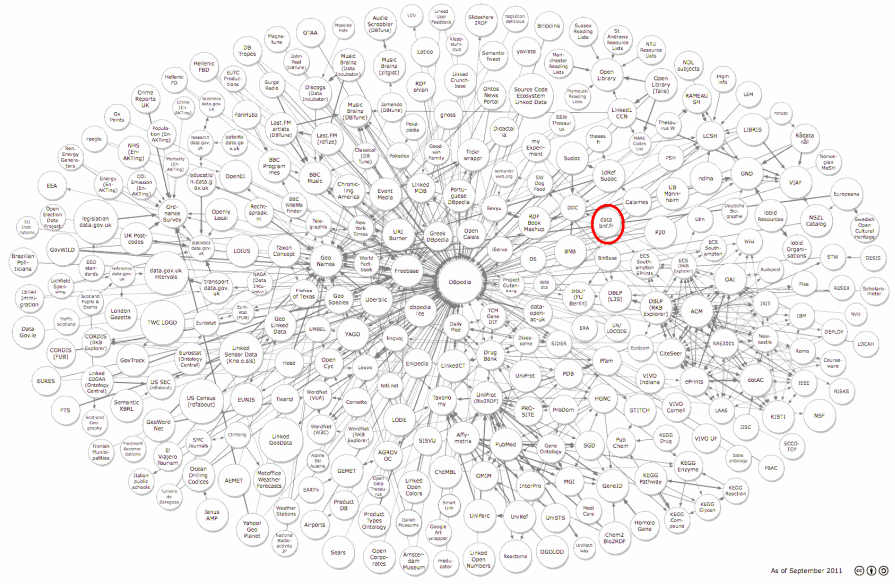
- **keep permanent URIs**, which also have to be understandable for the user, to refer to resources with useful information, and be integrated in a graph;
- build a content **negotiation system**;
- **use an RDF-compliant data model**, with standard vocabularies (basically SKOS, RDA and FOAF);
- use **existing vocabularies** as long as possible;
- use a **specific vocabulary** only for **classes and objects that are specific to the library;**
- **align our data with external data, from** the Library of Congress, the Deutsche Nationalbibliothek, Geonames, the Thésaurus W, for instance.

[4] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[5] http://www.lcweb.loc.gov/ead/
[6] http://dublincore.org/
[7] http://gallica.bnf.fr/
[8] http://expositions.bnf.fr/
[9] W3C Incubator Group Report. Library Linked Data Incubator Group Final Report. 25 October 2011.
URL: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/



How data.bnf.fr works

Trying to provide a significant hub in the "linked data" cloud is also a strong strategic issue, mainly in terms of bibliographic descriptions: choices concerning the FRBR model, Resource Description and Access (RDA), and the integration of EAD. This approach is somewhat innovative in libraries. Though they undoubtedly have a strong tradition of standardization, around interoperability and efficiency for sharing data, knowing how library data can become a significant hub in the Linked Data cloud is new issue. Besides, the **legal issues**, around dissemination and value are important. Data.bnf.fr is an open data project: RDF data can be freely retrieved and reused by anyone, according to our user licence.



Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

We endeavour to build a website based on a "semantic web" friendly data model.
The BnF has chosen to make its RDF data freely available and reusable.

→ Embedded data in the html pages

In our attempt to make the pages as useful as possible, we extract the most relevant descriptive data to integrate them into the source code of the HTML web pages as RDFa markups. Moreover we use the shared vocabulary from **Schema.org** to mark our pages with microdata, which is very important for many search engines. It is of course a way to make these representations of real-world things easier to index, with efficient and relevant computer-readable elements. We also put markups from the **Opengraph Protocol (OG).**

We tend to believe that this "open data" approach has to keep connected with the work on Web pages: not only because the data modelling around concepts leads to gathering relevant information around a single URI, but also because one of the basic "semantic Web" rules is to provide useful and human-readable information from these URIs.

The Web pages are structures with embedded data.
The type and formats of our RDF data are similar to the one available on the Open data cloud.
These two ways (RDF and HTML) to access data are really complementary.

Conclusion

data.bnf.fr does not replace existing catalogues: it is only a way to help users find what they need in our resources, to display content, links, services and share them more easily with others. This data-oriented approach of the Web interface enables us to combine our traditional missions with state-of-the-art researches and evolutions.

data.bnf.fr is still under development. This first version of the website holds the main French literature authors. The number of pages will be increased with new works (musical works for instance) and authors (composers, lawyers, authors from the Antiquity…) and "subject" pages will be created.

In the long term, we would like to link data to other cultural institutions, such as universities, archives and museums. In France, several "semantic web" projects are under development. Many difficulties have to be overcome, though, because of the differences concerning the kind of resources and of data.
Since many cultural institutions are now facing the same issues, in terms of Web presence and work on descriptive data and metadata, we hope it will be an opportunity to share experiences, in a more collaborative way.

Contact: data@bnf.fr
15/11/2011