



Gallica 2.0 : a second life for the Bibliothèque nationale de France digital library

Catherine Lupovici, Noémie Lesquins

Meeting: 146 Information Technology
Simultaneous Interpretation: No

WORLD LIBRARY AND INFORMATION CONGRESS: 73RD IFLA GENERAL CONFERENCE AND COUNCIL
19-23 August 2007, Durban, South Africa
<http://www.ifla.org/iv/ifla73/index.htm>

ABSTRACT

Gallica, the online digital library of BnF has been created as a pilot on the Web in October 1997, and the current version has been opened in January 2000 corresponding to the web technologies of that time. BnF took the opportunity of the work requested by the French Government to propose new services as a contribution to a European Digital Library and to rethink the way technologies and contents can be used to offer a more user centric digital library. A mock-up was designed as a prove of concepts building on the web 2.0 technologies and on the knowledge BnF already gathered about users requirements and behaviour through studies like BibUsages. The work achieved is intended to provide an intuitive and progressive appropriation of the services from discover, to search and finally contribution to users communities. The discovery functions are offering facets and thematic interfaces. The mock-up was used for users study through focus groups and the kernel functions implemented in a living prototype opened at the French book fair in Paris, end of March 2007. A new version of Gallica is being prepared, that will include at the end a semantic search engine. All the advanced functions and services are offered in a way that can be diverted by the users to create new usages.

BnF online digital library context

Gallica the BnF's online digital library is available on the Internet since October 1997 as a Web application. A pilot version with a sample of digitized collection was first open, and then the full collections were offered in January 2000 building on the pilot phase results and early users' reactions and feedback. The same version is still available and it will be progressively replaced by the new application under development, starting first with the printed material surrogates.

Gallica users are both registered users connected from inside the Library and worldwide users. The service is available 24 h and 7 days and offers printed material of different types such as books, serial publications, newspapers, but also manuscripts, prints, photographs, maps and a small sample of sound.

As for a physical library, the usages are related to the contents and the services offered. In a Digital library they are extended to new users by the technical ability to be indexed by general search engines as well as by more specialised other online services.

Gallica 2000 contents and interface

BnF started its digitization program in 1992 by printed material with the objective to offer a francophone encyclopaedic library on line. The collection policy has always been rather oriented toward scholars.

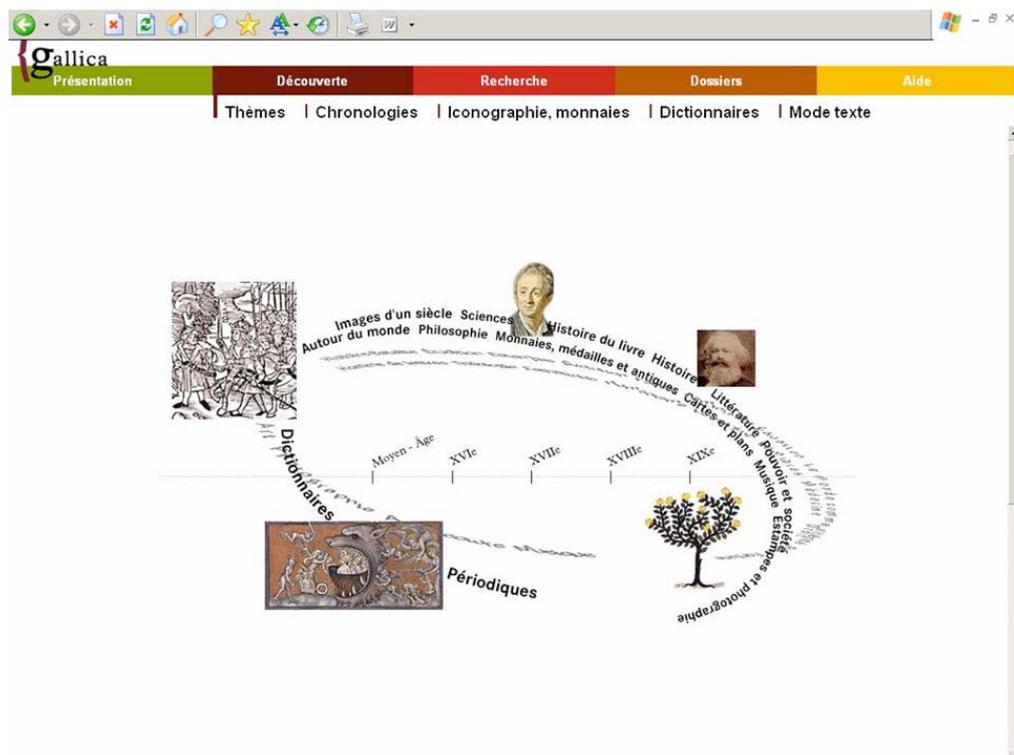
The printed material has been reformatted up to 2006 in image format for the whole document, complemented by a table of contents in text format allowing for some textual search inside the publications. Few documents are in full structured text. The newspapers program was initiated in 2004 and the documents are reformatted in image and text provided through rough OCR.

The digitized surrogates are described in the general library catalogue and displayed in the OPAC in the same way as the other surrogates like microfilms. They are directly available through a click.

The Gallica interface offers two main access modes: classical search and discovery.

The search interface offers a rather simple search compared to the OPAC interface. You can search by author, title, and subject. A full text search facility allows searching the bibliographic description, the tables of contents, and the captions of the images (which are not included in the MARC bibliographic description).

The discovery interface has an encyclopaedia structure in a hierarchy of themes and a chronological structure. It also includes a specific entry for the reference documents (bibliographies, dictionaries, biographies) and a specific one for the serial publications. In addition specific thematic corpus are offering zoom on the digital collections through folders, like permanent exhibitions of a subset of documents. This encyclopaedia approach was launched in fall 1999, and the granularity of this discovery mode is in between a classical library classification and the Wikipedia. Articles are presenting the themes with links to recommended documents.



Gallica has been early indexed by the general search engines like Google, Yahoo! or MSN. In order to facilitate its reuse by service providers the descriptive metadata are now exposed for harvesting on a publicly available OAI-PMH server¹.

Gallica users studies

Since its opening in October 1997, Gallica's usage evolution and users reactions have always been carefully followed, combining different methodologies like:

- Continuous recording of statistics on number of visits and visitors as well as page-view statistics and number of different documents used.
- Periodic online questionnaires in order to qualify the sociological profile of Gallica users.
- A qualitative study based on a panel of users was conducted in the context of a Dissertation in sociology on the usage of electronic libraries.

This qualitative study was made in 2002² and it provided a very in depth knowledge of the publics and their usage of Gallica. The public of Gallica is composed of 75% of scholars. 60 % of the users are located in France and 40% abroad. More than 85% of the usage is made from workstations located outside the library. 75% of the users enter into the Gallica application from a non BnF site (search engines, Internet provider, and other sites with links to Gallica).

This qualitative study was carried out through the installation on the end users personal computers at home of a software recording the whole activity on the Internet during 6 months. The data recorded demonstrated how Gallica usage is integrated in a general practice of the Internet and what specificities the usage of an online library has. Gallica users are big users of the Internet, with a general practice of information retrieval, and an important usage of cultural portals. Gallica sessions are longer than the average of sessions on the Internet and they are less alternate with the usage of other sites than the other sites used. Nevertheless the use of Gallica happens in the completely open context of the global Internet. The user is not anymore prisoner of the physical close context of the library. The tools used are not limited to the tools offered by Gallica itself and for instance the general search engines are used to search the collection contents. Gallica also opens on the external world and the study showed a link between non commercial and commercial services in the real practice of the users.

Gallica 2.0

BnF took the opportunity of the work requested by the French Government mid 2006 to propose new services as a contribution to a European Digital Library and to rethink the way technologies and contents can be used to offer a more user centric digital library. A mock-up was designed as a prove of concepts building both on semantic Web and on the Web 2.0 technologies as well as on the knowledge BnF already gathered about users requirements and behaviour through the previous studies made on Gallica.

The work achieved is intended to provide and test an intuitive and progressive appropriation of the services from discovery, to search and finally contribution to any users' communities. The mock-up was used for users study through focus groups and the kernel functions implemented in a living prototype opened at the French book fair in Paris, end of March 2007. An online questionnaire completed by the e-mail received has been used to assess the functions and interface provided in Europeana.

A new version of Gallica is being prepared, that will include at the end a semantic search engine. A specific additional work is also done on the Web 2.0 aspects which were not explored enough in depth during the Europeana project.

¹ *Le serveur OAI de la Bibliothèque nationale de France* <http://bibnum.bnf.fr/oai/index.html>

² *Users and Uses of Online Digital Libraries in France*. ECDL 2003.
http://bibnum.bnf.fr/usages/bibusages_ecdl2003.pdf

The new functions and interface

Europeana's mock up (<http://maquette.bnf.fr/labs/VersionAnglaise/scenario/Europeana.html>) is a succession of html pages with specific links illustrating all functions and describing a progressive use of all services by a user.

The five main use steps are: discover and browse, search, view, work and share.

Each step relates to specific screens and tools:

- Discover, browse and search are related to the home page and search results pages, with a blue graphic identity, and tools combining search engine performances (one-box search, refining or enlarging the search according to semantic analysis) and library catalogues (advanced or faceted search, results sets, refining according to metadata).
- View is related to the material visualization page, with an orange graphic identity, and to tools for navigation, browsing, searching into the item, downloading, printing, saving its URL for further use.
- Work is related to all personalizing features, graphically designed in green. The range of services goes from selecting and storing items into a personal private space named "My Library" to annotating and indexing each of them, as well as personalizing interfaces and services.
- Share refers to collaborative services, graphically designed in purple, offering each user the opportunity to create or join one or several communities ("Groups") in order to select, use and analyse specific material, to create and share new access points and perspectives on the collections.

Tools and services related to those five major uses are disseminated throughout the Digital Library. Three main pages gather and illustrate all functions described:

- Europeana homepage opens to the Digital Library collection through a rich variety of points of view and access, combining Web and catalogue performances, library and users communities approaches.
- Europeana search results page offers various features to refine or expand a search both based on semantic analysis of the request and user generated collaborative content.
- Europeana viewing page provides the user with a wide range of tools and services for a comfortable and in-depth navigation in the digitized item, for personal annotations and tagging and for sharing information and work with other users.

The Web and the Catalogue: browsing the collection

Large information spaces such as a Digital Library of millions of items of various kind, date, language, subject etc. can be quite confusing.

Browsing interfaces based on hierarchical faceted metadata are a way to guide the user toward possible choices and to organize the results in order to refine or expand the search.

Thematic search is based here on the 2 first levels of Dewey Classification, mainly used in Libraries around the World.

The Web and the Catalogue: full-text search

According to nowadays Web use, a one-box search allows to search one or several terms on an entire collection metadata and full-text content.

The advanced search form accessible in one click combines Boolean and proximity operators, catalogue fields and filters based on metadata (language, type of document, date etc.)

Full-text search, simple or advanced, is available on all pages.



The Library and the Community: social tagging and networking

In the recent Web evolution, many Websites use folksonomies to describe, label and retrieve their content.

Social tagging offers through user generated indexing and classification a more flexible and a definitely user-oriented view on the collection, allows unprecedented connections between items and enrich related content browsing possibilities.

Social networking around the Digital Library already exists for specific subjects such as Genealogy or History of Science.

Web communities can gather their own collection among the Library's offer in order to share comments, tags, discussion and generated specific views.

Semantic Web and the metadata: refining and expanding the search

For a flexible and flowing browsing, links based on faceted metadata allow users to refine and expand the current request, by adding or deleting selected terms, while visualizing simultaneously the generated results and keeping a representation of the whole collection.

This use of metadata is combined with a full-text search and a semantic analysis of keywords based on thesauri or ontologies, relating items with each other, either to refine or to expand the search.

Europeana search results page: digitized items networks in a semantic and collaborative Web environment

Semantic Web and the metadata: graphical visualisation

Semantic Web technologies allow to group metadata and show connections between items or concepts. An alternative to traditional record display, graphical visualization of search results would the users to browse and move visually within the collection.

The screenshot shows the Europeana search results page for the query 'newton'. The interface includes a search bar at the top with the text 'newton' and an 'Advanced search' button. Below the search bar is a 'Refine' sidebar with several filter categories: 'Your selection' (Newton), 'Media', 'Language' (English: 70, French: 48), 'Author' (Newton, Isaac: 49; Flammarion, Camille: 28; Arago, François: 10), 'Subject' (Science: 602; Mathematics: 201; History of France: 53), and 'Keywords' (Newton's Polygons: 15; Telescope: 3; Galileo: 8). The main results area displays a list of search results, including 'Newton's Opticks. Tome 1 / Isaac Newton, ... 1787', 'Newton's Opticks. Tome 2 / Isaac Newton, ... 1787', 'Philosophiae Naturalis Principia Mathematica, Tome 1 / Isaac Newton, ... 1759', 'Philosophiae Naturalis Principia Mathematica, Tome 2 / Isaac Newton, ... 1759', 'The Method of Fluxions and Infinite Series / Isaac Newton, ... 1740', 'Les fondateurs de l'astronomie moderne / Joseph Bertrand, ... 1865', and 'The History of Astronomy in the 18th Century / Delambre, Jean-Baptiste, ... 1827'. On the right side, there are several panels: 'My library', 'Sort' (with 'Relevance' selected), 'Tags' (showing 213 materials tagged with 'Newton'), 'Groups' (Astronomy for kids, Histoire des sciences), and 'Related resources' (Search in: BN-OPALE+, CCFr, TEL, voila.fr; Related websites: AstroWeb, Astrophysics Data System, The International Astronomical Union).

Semantic Web and the metadata: sorting results

Sorting by metadata is a classical way of organising results in a catalogue according to the user's centre of interest: author, title, date, etc. Search engines and semantic tools offer a wider choice: Relevance is based on objective and transparent criteria: FRBRization, place and number of the searched terms in the content, proximity for a multiple terms search, lingual relevance based the language of the interface etc.

Popularity emphasizes on the most popular items based on use statistics: the most viewed, the most tagged, the most recorded for a personal use, the most used and commented by communities. Result sets rely on the most relevant criteria to group results in 3 or 4 lists.

The Library in the Web: expanding the search to related Web resources

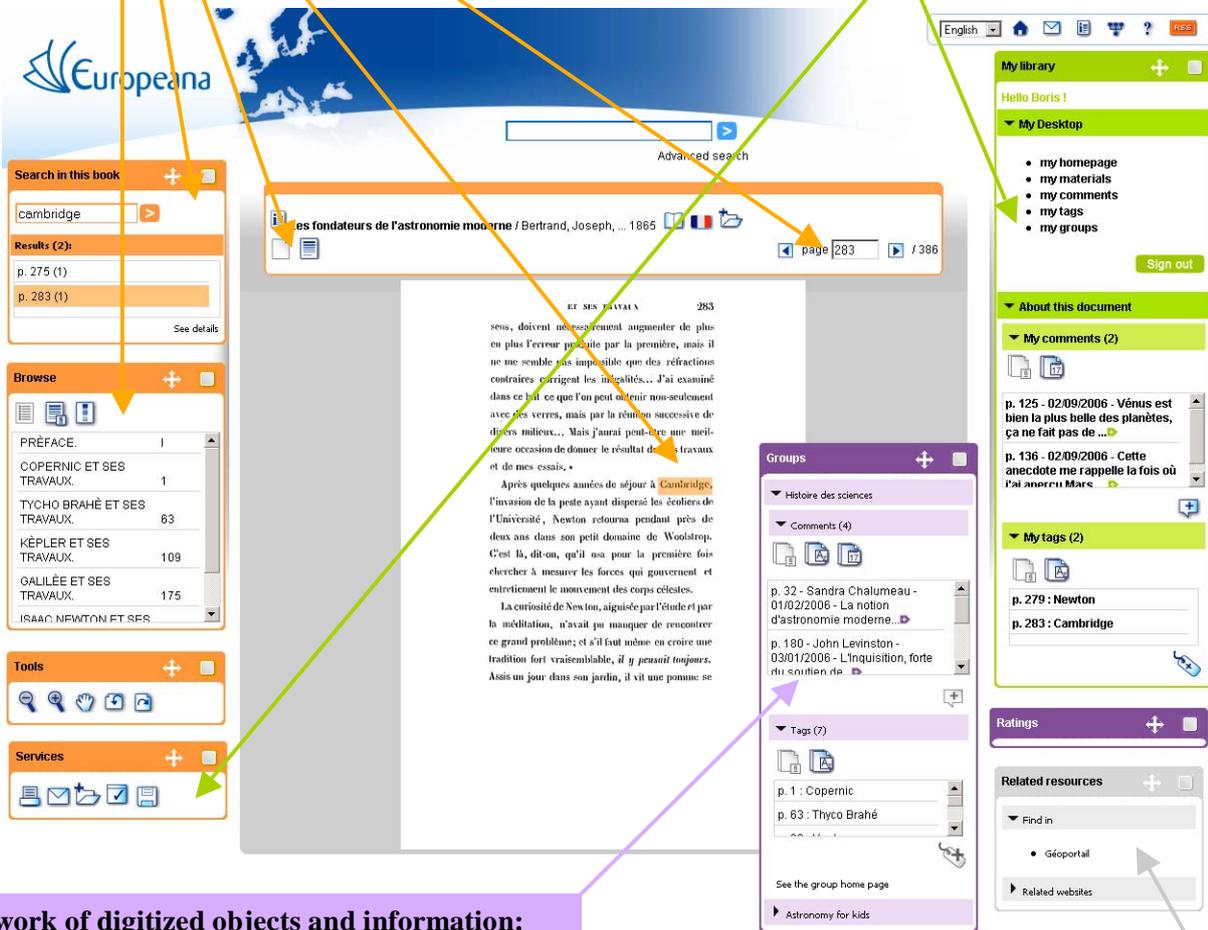
User surveys state that the use of Digital Libraries is often made in connection with other Websites (search engines, online bookstores, libraries catalogues etc.) A librarians' selection of search tools and valuable Websites in relation with the request provides the users with a choice of resources.

Browse and search in the digitized object

A wide range of tools provide the users with a comfortable navigation in the digitized object: XML tables of content, thumbnails and linked page numbers are complementary to page to page browsing, while full-text search allow specific search within a text with a simple XML alto structure. Each page can be viewed both in image and in text modes.

Personal work on the digitized object

Downloading on local disk is a regular use of Digital Libraries, as well as printing. Yet online reading is at stake, especially for regular users. A permanent URL for each page of each item provides the users with several ways to keep track of a digitized object: book marking in one's Web browser or one's email box, or storing in one's own Digital Library. Storing and organizing one's own online collection, indexing and annotating it with one's own taxonomy and point of view allow a real appropriation of digitized objects, easy to retrieve and enrich anytime.



A network of digitized objects and information: collaborative work on the digitized object

Sharing personal annotations and tags and exchanging with fellow users is possible with a set of similar tools as for personal use. Users' communities can be private, restricted or public and are free to set their rules according to their scientific and content purposes. With moderation and control by the librarians, online collaboration contributes to improve access to the collection, by enriching perspectives, creating specific tools for specific needs on specific sets of objects, and generating networks of digitized objects and information.

A network of digitized objects and information: expanding the use to related Web resources

Text structuration, semantic analysis, librarians' reference work and Web services help connecting the digitized object with other information and tools on the Web.

User surveys on Europeana

Several user surveys were lead on Europeana mock-up and prototype: Focus group discussions, questionnaire online, marked pages, free-text reactions and suggestions. About 550 users gave their opinion about Europeana access, services and interface, helping us to get to know better the potential public and uses of a second generation Digital Library.

In terms of public profiles, in comparison with Gallica's users known after a study led in 2002, we observe the public is slightly younger, even if retired people are a stable part of the users. 40% of the identified and faithful users are library or education professionals, or by any means with a high education level. Research on the Web for those specific users implies the use of general Websites (search engines, online shopping) but also with other digitized books sites (Google Book Search, Gallica). The main research scope is genealogy or humanities. Many Internet and Information system professionals also showed some interest in the prototype.

All services provided in the mock-up and the prototype were considered useful by all users. Yet there is a large demand for a better quality (better search engine relevance, better OCR or downloading performance) or for more services (data exporting services, translation services, and print-on-demand services). The idea of integrating users' communities in the Digital Library is being very successful among all kind of public, in accordance with the worldwide success of the online collaborative encyclopaedia Wikipedia. Collaboration online opens to a wide range of subjects and uses (improving OCR, summarizing books, educational purposes etc.) but triggers also a strong demand for control and validation of user generated content. The library is seen and wanted as a space for leisure, learning, working, and sharing within a safe and structured environment controlled by acknowledged authorities.

In terms of interfaces, ergonomics needs to be improved, for a better accessibility (compatibility with all terminals and browsers, accessibility for the impaired, simplicity of use for everyone) and a more comfortable online reading and browsing.

Trust and fluidity of use are the two major expectations for the Digital Library. But it is important also to keep in mind a need for the librarians, the collection and the users to meet not only within the Library but also in the Web flow, for instance through search engines or RSS feeds.

Gallica and the Web 2.0: lessons learned and future directions

The Web 2.0 is a concept defined by Tim O'Reilly to characterize the new generation of web players arising after the bursting of the .com bubble in the fall of 2001³. This marketing concept wraps together several principles out of which the following points are important for the design of a user oriented digital library:

- On the technological aspects it means new tools and new services
- On the sociological aspect it means new usages and new user behaviour

The modernization of a Digital Library like Gallica, which was built in the Web 1.0 technological and societal environment, must integrate those two points.

On the technical aspect the Digital Library must be open and exploitable by external new tools as it was already mandatory for the first generation tools. It is very important for the contents to be present in the Web workflow in order to be discovered when using the new dissemination tools (RSS, wikis, and blogs), the new research tools (human search engines, collaborative search engines, and mashups), the new social tools for sharing and contributing (social bookmarks). It is mandatory to catch the attention of users adding on the contents (clustering, recommendation ...)

³ *What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software*, by Tim O'Reilly. 09/03/2005. <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>

On the sociological side, the users are taking over the previous function of institutions when they are becoming authors and publishers and in some way librarians through the usage of blogs and through the contribution to wikis for instance. The profusion of the contents modifies the previous strong attention paid to a limited number of resources and from now the users are more interested in the global network whose they are becoming another element. The integration of the users' contribution is becoming a key point through tags, comments, blogs, and wikis.

The new practices are characterised by a change of the authority of the institution into the popularity of the contents and services, as well as a change of the relevance of a search result into the influence of the users who are recommending, commenting and tagging contents

The issue for a Digital Library integrated into the Web 2.0 is to define which services to add on its traditional offer as a normal extension of library services and which ones have to remain external services building on the contents the Library exposes on the Web.

After the exploratory phase of the design and assessment of the Europeana prototype, the points which are still discussed for Gallica 2.0 are the sociological aspects and the users' contribution. Which importance to give to popularity and users influence aside the traditional functions of authority and relevance? The user surveys gave a strong requirement for such services in addition of the continuation of the more traditional ones. They are attached to the authority represented by the library. A specific working group on user collaboration has been set up to define the strategy for the implementation of user contribution into Gallica 2.0. One of the first results is that such an evolution is also a requirement for the public services in the physical reading rooms as it is a larger social phenomena then just the Web. This working group will also look at the legal aspects of the institution's responsibility about the contents created by the users.