Submitted on: March 15, 2013

Customized OPACs on the Semantic Web: the OpenCat prototype

Patrick Le Bœuf
Bibliothèque nationale de France, Paris, France.
patrick.le-boeuf@bnf.fr

Copyright © 2013 by **Patrick Le Bœuf**. This work is made available under the terms of the Creative Commons Attribution 3.0 Unported License:
http://creativecommons.org/licenses/by/3.0/

Abstract:

OpenCat is a research and development project that was led in 2012 by the National Library of France (hereafter referred to as BnF), in partnership with the public library of Fresnes (a small town located in the suburbs of Paris), and the French software editor company named Logilab, which specializes in Semantic Web technologies.[1] The initial objective of this project was to enable a small library to take advantage of the publication of (part of) the content of BnF's catalogues as semi-FRBRized Linked Open Data on the data.bnf.fr Web site to enrich its own catalogue with the result of that semi-FRBRization process, along with snippets from external information resources. The outcome is a Semantic Web application that makes it possible for libraries to share the same dump of FRBRized Linked Data, while they can customize the way it is displayed to end-users and enriched with external information.
This paper will focus first on the data.bnf.fr project, which occasioned the OpenCat project and formed the basis for it; then CubicWeb, the technical platform that served to build the OpenCat prototype, will be briefly introduced; some features of the OpenCat prototype will be described; and lastly, some responses to the prototype from a panel of end-users will be analysed.

Keywords: Bibliographic data as Linked Open Data, OPACs and Library Linked Data, Data aggregation

1 THE DATA.BNF.FR WEB SITE, OR, FRBR IN THE LINKED OPEN DATA UNIVERSE

In the summer of 2011, BnF launched a new Web site, named data.bnf.fr.[2] This Web site, which was awarded the Stanford Prize for Innovation in Research Libraries (SPIRL) in

[1] For more information on the three partners in this project, see, respectively: http://www.bnf.fr, http://bm.fresnes94.fr/, and http://www.logilab.fr/. This project was supported by the French Ministère de la Culture et de la Communication.

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

2013,[3] was the outcome of over one year of efforts to find ways to make BnF's resources more visible on the Web and to transform the metadata that describes them into Linked Open Data available for anyone to reuse.[4]

BnF has two catalogues, *BnF catalogue général* for published resources (http://catalogue.bnf.fr), and *BnF archives et manuscrits* for archival collections and unique manuscripts (http://archivesetmanuscrits.bnf.fr), plus several databases that describe particular types of items or that focus on particular features of items otherwise described at a lesser level of granularity in either of the two main catalogues. In addition to that, BnF also has a digital library, Gallica (http://gallica.bnf.fr), which derives its descriptive metadata from the two main catalogues, and Web pages that correspond to BnF's activities in the field of cultural mediation, such as 'virtual exhibitions' (http://expositions.bnf.fr) or recorded lectures (http://www.bnf.fr/fr/evenements_et_culture/conferences_en_ligne.html). All these tools, although they share the same general objective of enabling end-users to discover BnF's resources, have separate interfaces, as they are built on different formats: *BnF catalogue général* uses a specific declination of the MARC family of formats, named INTERMARC; *BnF archives et manuscrits* is produced using the EAD DTD; and the metadata that serves to retrieve digitized items on Gallica is expressed in Dublin Core. The data.bnf.fr project originated from the urgent need to propose a common interface allowing end-users to query all these information repositories at once, while preserving the distinct structure, granularity, and ontological commitment of each of them individually.

The data.bnf.fr Web site (http://data.bnf.fr) consists of human readable pages, the content of which is also available as RDF and JSON declarations in order to be exploited by Semantic Web applications. The objective is not to *substitute* this Web site to the dedicated interfaces of the BnF's existing catalogues, but to provide Web users who are not accustomed to searching the BnF's resources with a pivot interface that digs such resources out from the deep Web (hence the need for human readable pages), in addition to making bibliographic information available as Linked Open Data (hence the need for RDF and JSON declarations). The RDF and JSON declarations are provided under the French Government's open license, which is very similar to the Creative Commons CC BY license. Indeed, the content of the data.bnf.fr database represents the first and, to this day, most significant data set that was made available from the French platform for public open data, data.gouv.fr (http://www.data.gouv.fr/DataSet/30383137).

As of 1st July 2013, the data.bnf.fr site comprised over 600,000 pages, divided up into three main categories: Authors, Works, and Topics.[5] Each page is devoted to a single author, work, or topic, and includes brief citations of bibliographic records associated with that author, work, or topic; these brief citations are provided with links that lead Web users into the BnF's catalogues. 'Work' is not to be understood here in the sense Bibframe uses it, but in the FRBR sense of the term. This concept was chosen as a convenient collocating unit for bibliographic data, because the articulation between a cultural object (such as a novel, a

[3] For more information, see: http://library.stanford.edu/projects/stanford-prize-innovation-research-libraries-spirl/bibliothèque-nationale-de-france-judges.

[4] The conceptual model that lies at the heart of the data.bnf.fr database is available, in a simplified form, from: http://data.bnf.fr/semanticweb-en#Ancre3.

[5] Two further categories are represented: Years, and Places.

poem, a scientific study, a symphony or a motion picture), its variants (be they compositional, linguistic, performatory, etc.), and the many ways these variants can be packaged for dissemination within a society, was deemed to have a broader significance than just as a construct of the entity-relationship model designed by IFLA for bibliographic data in the 1990s. Indeed, we regard it as having been the implicit underlying conceptualization of the bibliographic universe for at least two centuries. We think that these distinctions lie at the heart of the definition of networks of cultural objects, even in the Semantic Web era. Unlike some 'FRBR sceptics' who have recently appeared and who argue that FRBR is not compatible with the very spirit of Linked Open Data, we believe that the fact that not all other communities outside libraries are aware of the FRBR model or willing to implement it should not refrain librarians from using it to inform the underlying structure of the Linked Open Data they produce. We feel encouraged in this direction by the existence of Wikipedia pages devoted to works, which proves that even in the 21st century it still makes sense to organize a cultural discourse around the notion of works, rather than to content oneself with listing myriads of publications defined on the basis of the ISBN criterion only. We are also convinced that the FRBR distinction between Work, Expression, and Manifestation can actually facilitate, rather than hinder, the linking functionality within the Web of Data, as it allows other communities to reuse indifferently either the complete RDF graph (from the Manifestation level to the Work level) that describes a given publication, or just that part of it that relates to the specific level they may be interested in and find relevant for their own needs, e.g., just the information that pertains to the Work and Expression levels. From a pragmatic point of view, taking the FRBR Work entity as its collocating unit for exposing bibliographic data as Linked Open Data also enables the data.bnf.fr Web site to not simply duplicate the catalogue and result in millions of atomic pages, but to organize bibliographic information in a way that makes sense for end-users because it corresponds to the way they envision cultural history.

At its current stage of development, however, the data.bnf.fr project does not propose a complete implementation of the FRBR model. It can be more aptly labelled a semi-FRBRized view of the BnF's catalogues than a fully FRBRized one. At the time being, due to the nature of the legacy data that has to be transformed into Linked Open Data, the Expression entity is not represented in the data.bnf.fr Web pages, but it is included in the plan for future development. As other experimentations in FRBRization, such as OCLC's FictionFinder, have amply demonstrated, the Expression entity, although it is the most 'natural' and self-evident in the FRBR model, is also the one that proves most difficult to be dug up from legacy data – an interesting issue from the point of view of the philosophy of cultural history.

The data.bnf.fr structure elaborates on the authority work already performed in the BnF's catalogues. Authority records are created at the BnF for all musical works that are available in notated form, all 'art music' (as opposed to both folk and popular music) works that are available as sound recordings, all literary works that are mentioned in the IFLA lists of 'anonymous classics',[6] and all works of any type whenever a controlled access point based on the uniform title that identifies them is needed for subject indexing. The data.bnf.fr 'Work' pages are automatically generated from those authority records, and bibliographic records that, in the catalogues, are not associated with them (most often because they describe publications of works for which an authority record is only usable as part of a subject access point) are aligned on them through specific algorithms. As of 31st January 2013, the

[6] http://www.ifla.org/node/4957

data.bnf.fr Web site contained links to 3.5 million bibliographic records, i.e., one fifth of the records present in *BnF catalogue général*. Over 70% of Web users who visit one of the data.bnf.fr Web pages follow the links to other Web pages created by the BnF (either the catalogues or the digital library), which is a good indication that the data.bnf.fr Web site is rather successful as a way of making library resources 'alluring' to Web users who were previously unaware of libraries' potential for their information needs. But the main objective pursued when publishing Linked Open Data is not just to have it read, but to have it reused by others who should feel free to do whatever they want with it. We are aware of several cases in which the content of the data.bnf.fr web pages was actually reused by developers who are totally independent from the BnF, e.g., the mobile application named CatBNF which makes it possible to browse the data.bnf.fr pages on an iPhone or an iPod, or the Web site named IF Verso, which is devoted to the notion of translation and proposes a catalogue of translations from the French based on the Work-centred structure of the data.bnf.fr Web site.[7] However, there may be other reuse cases of which we are not aware, as of course the difficulty with the freedom enabled by Linked Open Data is precisely that it is very difficult to keep track of all the transformations undergone by the bibliographic data thus exposed.

2 THE CUBICWEB APPLICATION FRAMEWORK, OR, THE SEMANTIC WEB SEEN AS A CONSTRUCTION GAME

The data.bnf.fr application was built using the application framework named CubicWeb, which has been developed by the French software editor company Logilab since 2001 (version 2 was released in 2006 and the current version, version 3, has been downloadable for free since 2008).[8] CubicWeb is devoted to Semantic Web techniques used as a means to federate data sets from heterogeneous sources and display them in a variety of ways corresponding to clients' various needs. CubicWeb was designed as a generic platform consisting of components that can be reused and rearranged in various combinations so that the time and cost of developing specific applications can be reduced. In CubicWeb's parlance, these components are called 'cubes,' because they consist of three elements: a data model that is expressed in the formalism of an entity-relationship schema, the logic that is required in order to manipulate the data, and the view code that makes it possible to visualize the federated data (user interface, export in various standards). The query language used by CubicWeb is RQL (Relation Query Language), is similar to W3C's query language SPARQL, and is closely related to the underlying data model.[9] This structure makes it possible to provide access to resources that are deemed useful for end-users, and that are identified by more or less perennial, dereferenceable URIs.

Since the data.bnf.fr application and Web site were developed by the Logilab company, and since the objective of the OpenCat project was to explore the possibility of federating bibliographic data from the data.bnf.fr pages, local holdings data, and other types of data from external sources in a single customized OPAC, it was quite logical for the Logilab company to propose itself as the third partner in that project.

[7] CatBNF: http://www.cnetfrance.fr/telecharger/catbnf-40306203s.htm; IF Verso: http://ifverso.com/.

[8] Documentation about CubicWeb is available in: *Logilab, CubicWeb – the Semantic Web is a construction game!* Last updated on Sep 18, 2012. http://docs.cubicweb.org/.

[9] For more information about the way CubicWeb is used for the data.bnf.fr project, see: Simon, Wenz, Michel, & Di Mascio (2013).

3 SOME FEATURES OF THE OPENCAT PROTOTYPE, OR, A NATIONAL LIBRARY'S OPEN DATA AT THE SERVICE OF A LOCAL LIBRARY

The initiative of proposing the OpenCat project emanated in November 2011 from the Fresnes municipal library. This initiative was quite consistent with a deeply rooted tradition there, as this library had been publicly advocating and internally implementing for a number of years a proactive policy aiming at reducing cataloguing time and costs by reusing both the authority and bibliographic records produced by BnF (Giappicioni 1998). The six objectives of the project were introduced in the library's proposal as follows: (1) facilitate the search for documents by displaying query results organized around the FRBR notion of 'Work', (2) use Semantic Web techniques to combine in a single hit list documents from various sources, (3) propose direct links to freely accessible digitized documents whenever possible, (4) enrich hit lists with contextual information taken from Open Data repositories, (5) propose a new approach to subject browsing through an intuitive graphic representation of the RAMEAU subject indexing language indexed as SKOS, and (6) use in the library world the same techniques and services as in commercial sites. This proposal was received favourably by BnF, as it was immediately regarded there as a convenient means 'to ensure that the way the BnF envisions the future of its metadata will ultimately meet practical needs and constraints from the national public library network rather than propose a scheme that nobody would use' (Illien 2012).

The OpenCat project was officially launched in May 2012, and was to last 15 months. By early November of 2012, Logilab already had delivered a first prototype. The current version of this prototype is publicly available from http://demo.cubicweb.org/opencatfresnes, with all the data that is relevant for the Fresnes library. Another experimental version is being tested. It contains data from other libraries in addition to Fresnes, displays extended functionalities, and has links to a larger array of external Web sites; however, access to that version is restricted through login and password. A password is given to any library that wishes to experiment the addition of local holdings information to the contents of OpenCat.

The first step consisted of developing a first data model and merging in one repository data sets from both BnF and the Fresnes library, plus data sets from external sources with the aim of enriching the catalogue. The data set from the Fresnes library did not comprise the entire catalogue of that library, but only that portion thereof corresponding to works present in the data.bnf.fr database. This portion represents about 4,000 bibliographic records out of the 230,000 that were copied in the data repository hosted by Logilab; these 4,000 bibliographic records describe manifestations of about 3,000 out of the 45,000 works to which a data.bnf.fr page is devoted.

The second step consisted of developing graphic mock-ups and a demonstrator that allowed the project team to perform tests on the available data, refine the data model, and select the navigation scenarios and functionalities that were deemed the most relevant among a number of proposed alternatives. What follows is a brief guided tour of that demonstrator.

The homepage of the public prototype consists of three sections: a) a Google-like search box in which users can enter any kind of terms, b) random proposals of digitized items, and c) an automatically produced timeline of works that were granted a literary award.

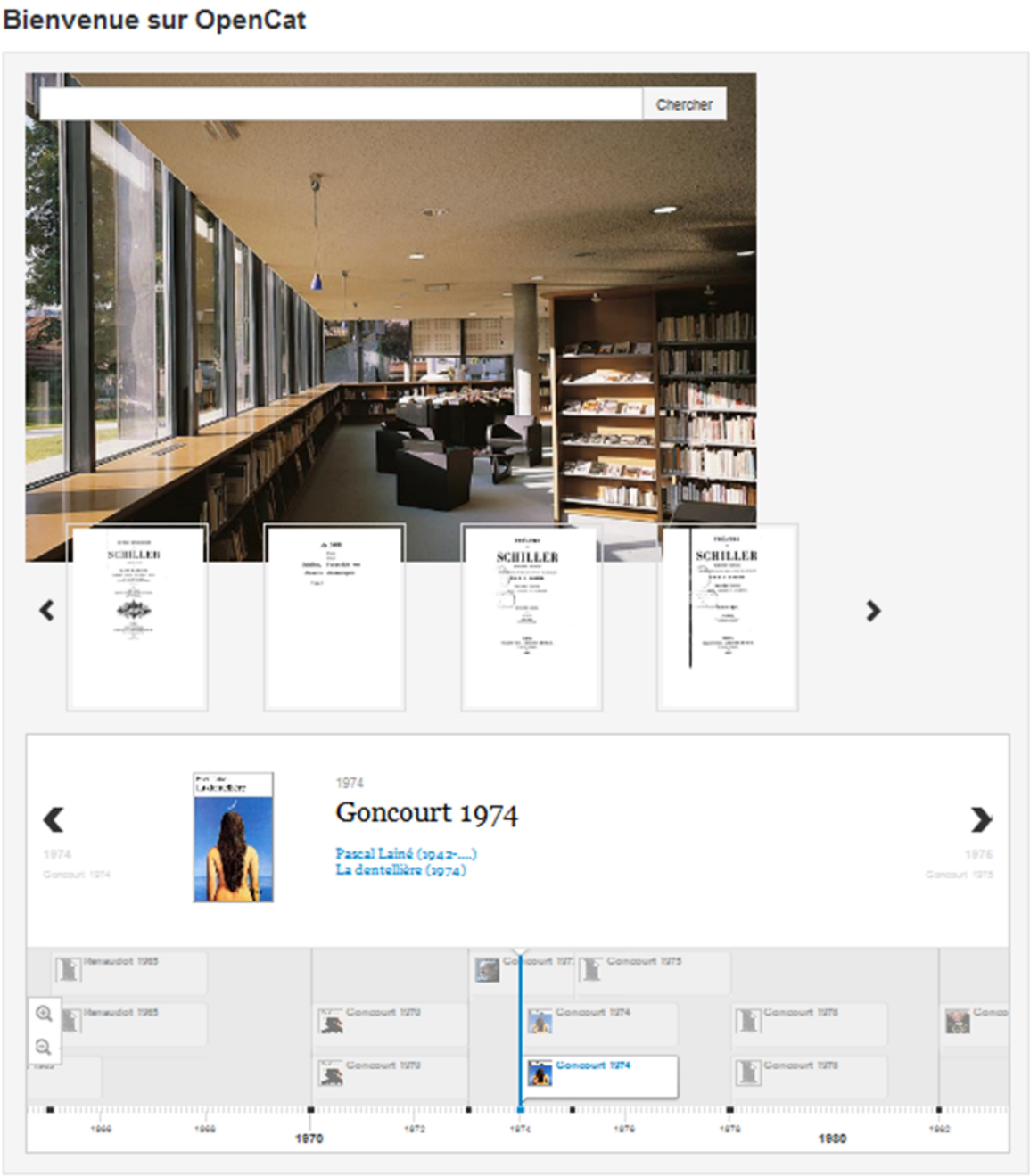

Figure 1: OpenCat Homepage

When users type the terms `Émile Zola` in the search box, a list of proposals is automatically displayed. This list includes the name of a composer who collaborated with Émile Zola, the title of a work by Émile Zola, and names of characters imagined by Émile Zola and for which a subject authority record exists, and the complete controlled access point for Émile Zola.

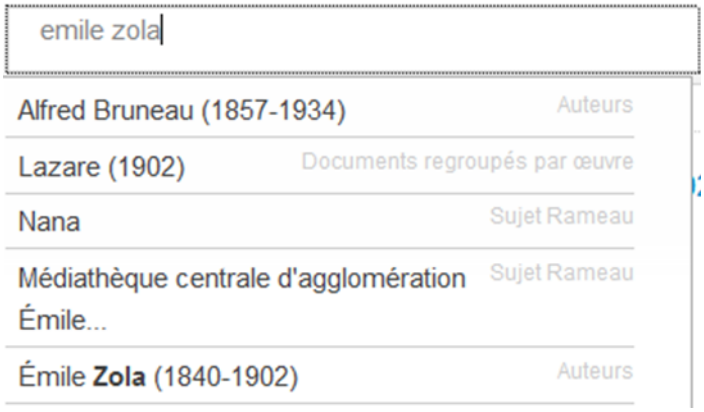

Figure 2: List of proposals for a query on `Emile Zola`

It is also possible to bypass the list of proposals, and launch the query without having to select one of the items in the list. In that case, the system returns a list of hits, with the possibility to refine the query, by picking a category of information, such as: person, corporate body, topic, title of work.

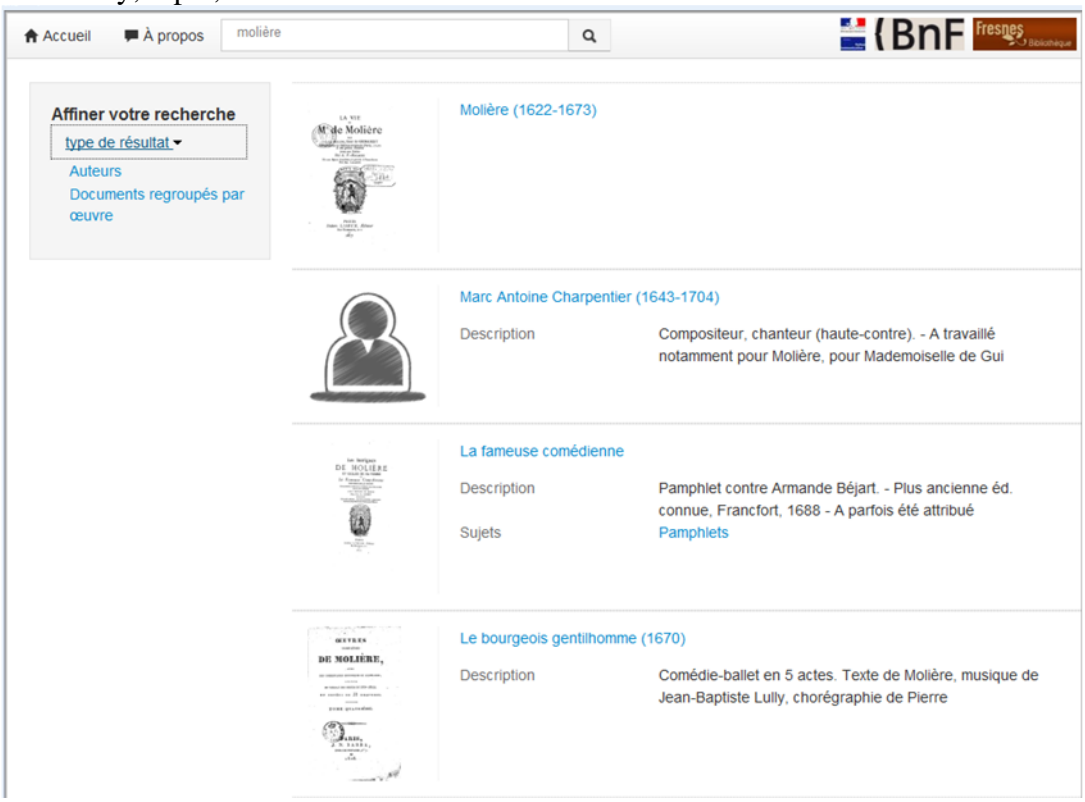

Figure 3: Refining a query after the initial list of proposals was bypassed

If a user looks for the name `Eugène Ionesco`, they will find some information elements from the authority record produced by the BnF, combined (without any manual intervention by the user) with a link to the data.bnf.fr page devoted to Ionesco, an excerpt from the biographic note devoted to Ionesco on the Web site of the Académie française, of which he was a member, and a link to the rest of the page from which that excerpt was taken. This biographic information is followed by three distinct sections: a) Ionesco's works, organized along a timeline according to the date of their first publication, b) Ionesco's works, organized as a list of clickable links to pages devoted to them, and c) links to 'other resources,' in this case a virtual exhibition prepared by the National Library of France.

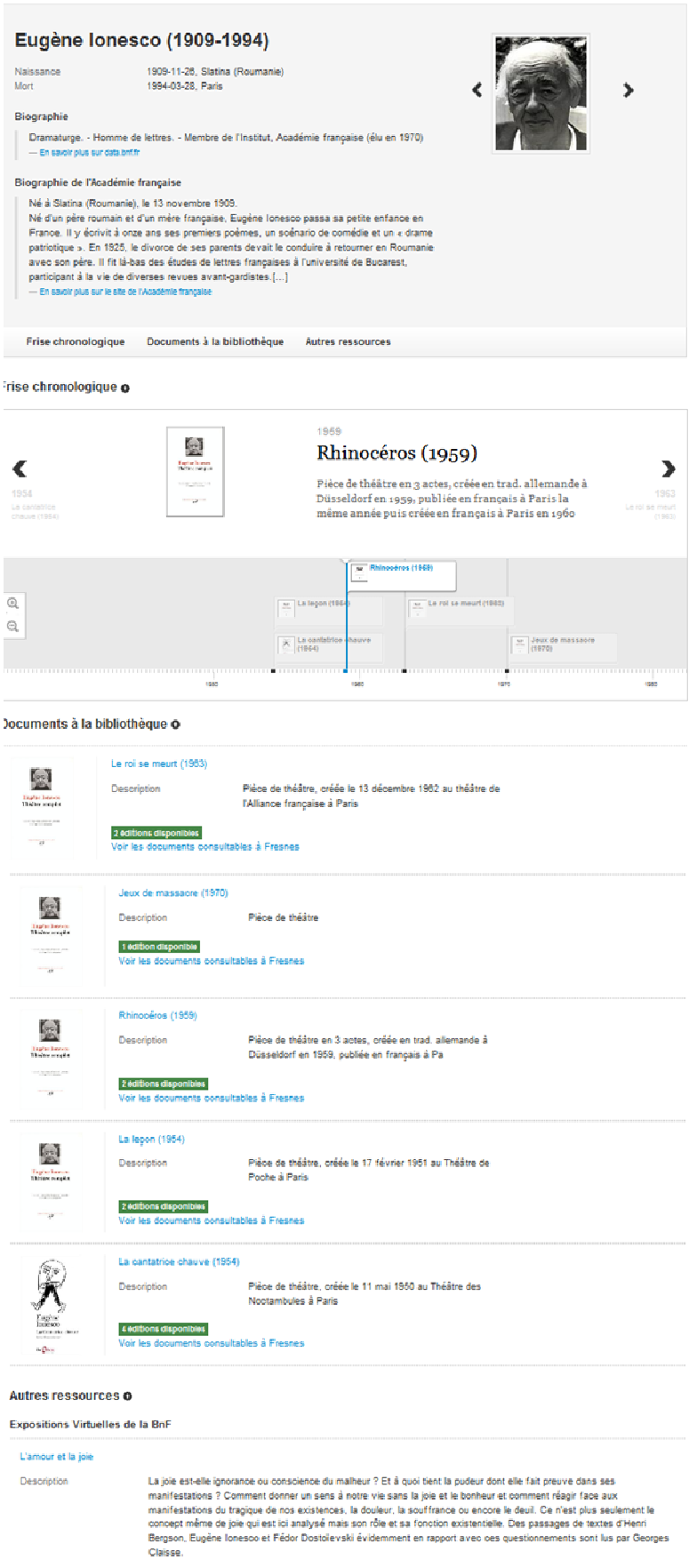

Figure 4: Author page for Eugène Ionesco in OpenCat

The experimental version of the prototype shows, in addition to all of the above, a link to the speech delivered by Ionesco when he joined the Académie française, links to pages devoted to Ionesco's contemporaries (in the future, it will be possible to replace this strictly chronological connection with more meaningful ones, such as: other authors who were active in the same field of interest, or who wrote works belonging to the same genre), and links to popular Web sites such as DBpedia, FreeBase, MusicBrainz or Flickr.

Figure 5: Author page for Eugène Ionesco in the experimental version

A user who types the term `Pathelin` in the search box will obtain a page devoted to the anonymous mediaeval work *Farce de maître Pierre Pathelin* that contains: a summary of the work, taken from a commercial Web site; a list of the seven editions of which the library holds a copy; and links to nine digitized items available from Gallica, the BnF's digital library.

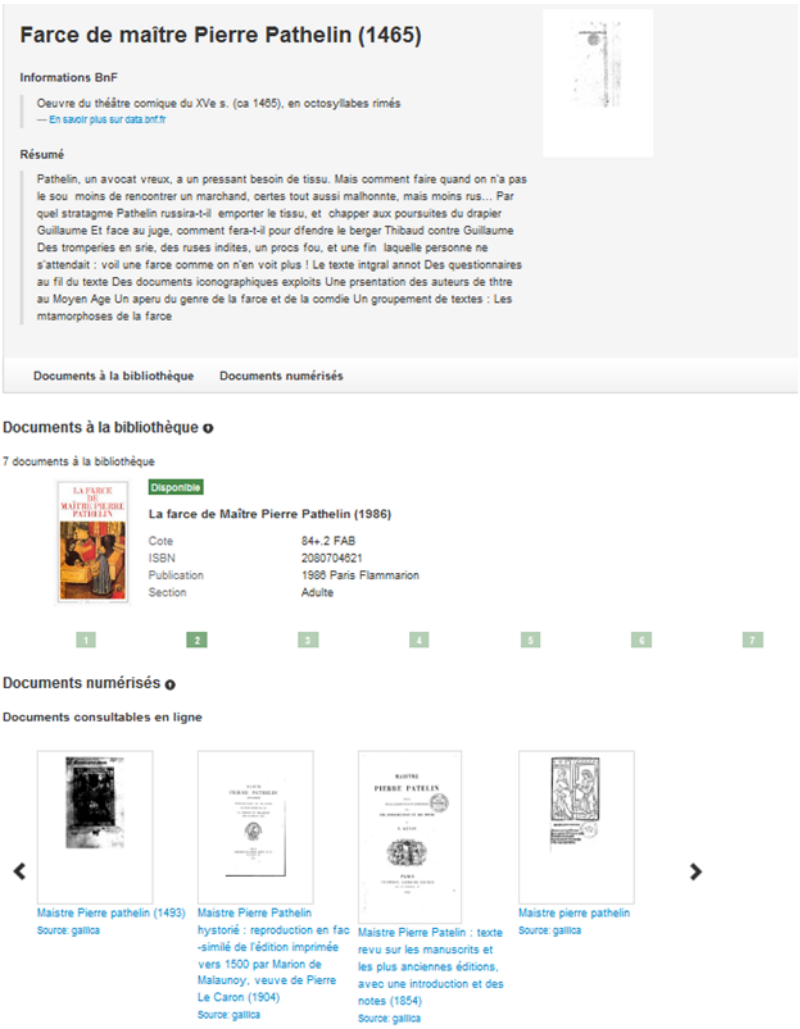

Figure 6: Work page for an anonymous work

If a user types the term `Hamlet` in the search box, the OpenCat demonstrator will display a list of various proposals, including: the play entitled *Hamlet*, the subject heading for the Ophelia character, and the subject heading for the Elsinore castle.

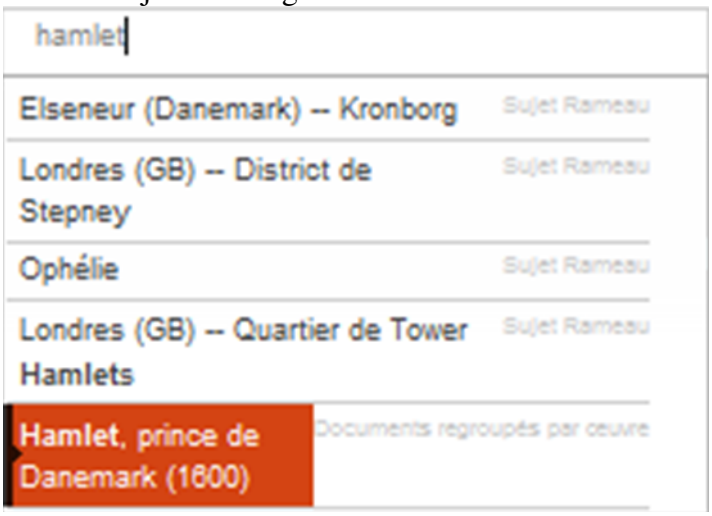

Figure 7: List of proposals for a query on `Hamlet`

If a user types the term `poésie`, the demonstrator will display a list of poets and poetic works from which to pick.

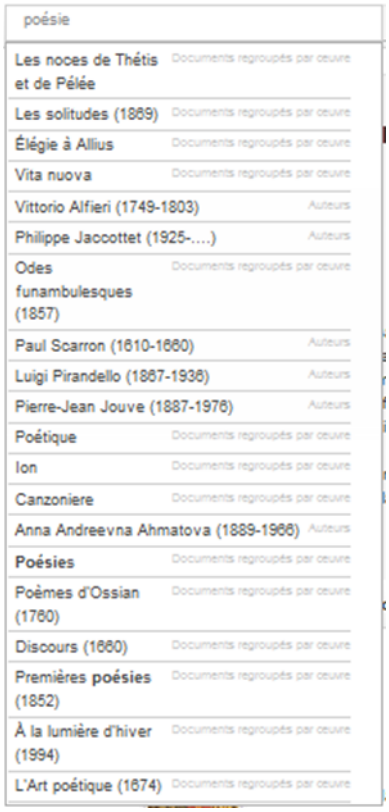

Figure 8: List of proposals for a query on `poésie` on OpenCat

The public version of the prototype does not allow one to search by subject, but the experimental one provides access to information taken from the subject authority file maintained by Centre national RAMEAU (hosted by the National Library of France and available from the data.bnf.fr site), links to associated bibliographic records, and links to external resources. For instance, a query on the term `Jeunesse` (youth) leads users to a number of recorded lectures available from Canal-U (an academic network Web site).

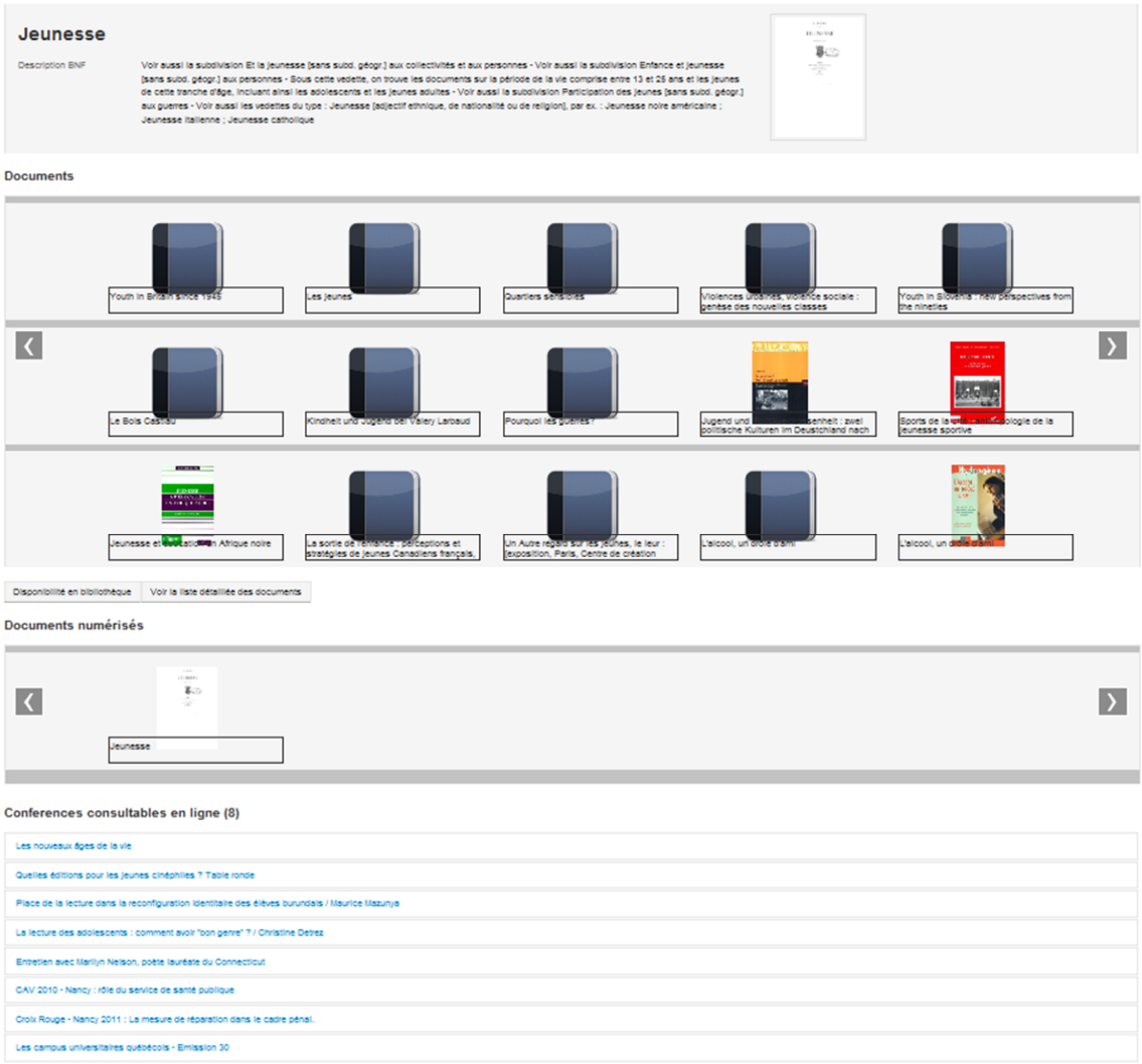

Figure 9: Returned results after a query on `jeunesse` in the experimental version

If a user types the ISBN `978-2-81000-220-7`, OpenCat will return the work *Madame Bovary*, an edition of which is identified by this ISBN, although the Fresnes library does not hold any copy of specifically that edition. This is made possible by the fact that, in the data.bnf.fr database, all editions of *Madame Bovary* are collocated on the page devoted to that work. It is therefore possible to imagine that the catalogue, in the future, will be able to notify users that the library holds no copy of the edition they searched under its ISBN, but that another edition of the same work is available to them.

Figure 10: Proposal for a query on an ISBN

The figure below indicates the sources from which information is aggregated in OpenCat.

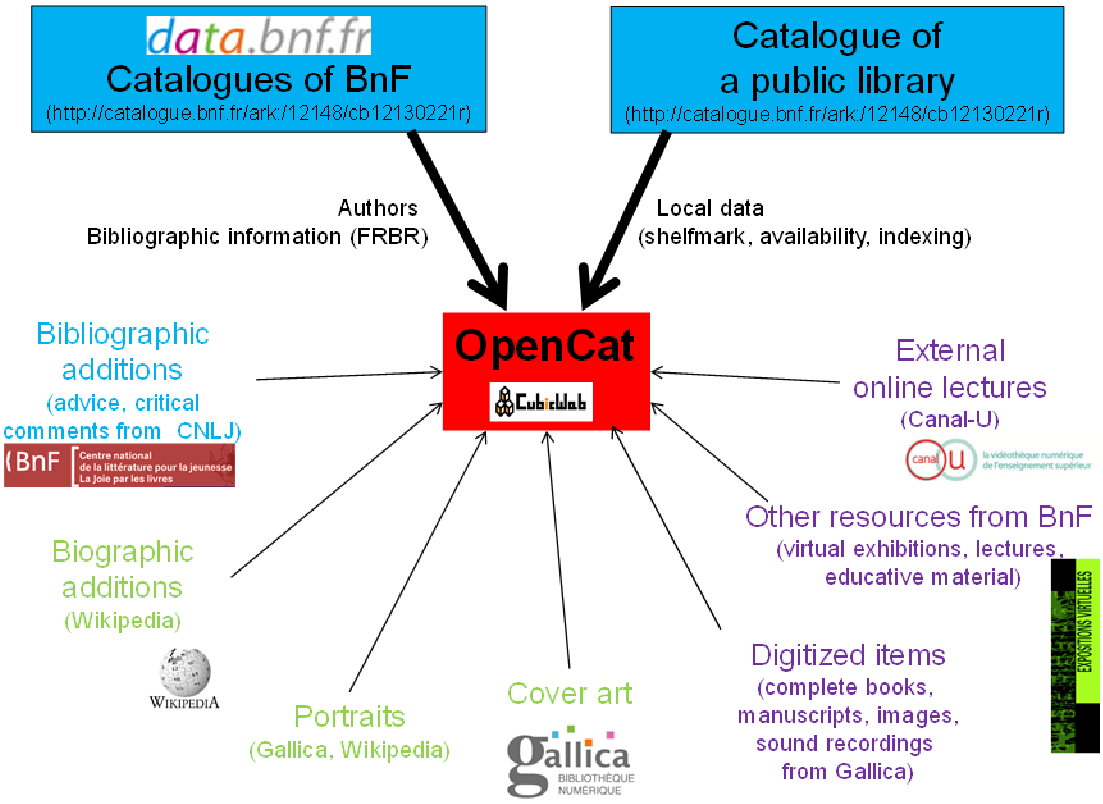

Figure 11: Sources for the information aggregated in OpenCat

4 SOME RESPONSES TO THE OPENCAT PROTOTYPE

In December 2012, ten patrons of the Fresnes library, aged between 34 and 63, were asked to test the OpenCat prototype. They were left free to discover it by themselves for fifteen minutes, after just a very brief presentation of its main functionalities. Then their opinions about that experimentation were collected by one member of the BnF staff and one member of the Fresnes library staff, who asked them to try a specific query on a predetermined topic that was known to highlight the main features of the prototype that it was important to assess. All panellists expressed their satisfaction that the traditional content of a library catalogue was enriched with information from external sources. However, this enrichment was also perceived as possibly blurring the identity and the very nature of the interface: some panellists felt that there was either too much information for a local library catalogue, or not enough of it for an online encyclopaedia. Inevitably, comparisons were drawn with Wikipedia.

On the whole, the panellists were satisfied with the ergonomics and ease of navigation of the prototype. However, the autocomplete function of the search box was perceived as too constraining and the order in which the proposed items are listed as illogical; none of the panellists seems to have discovered by themselves that it is actually possible to bypass that list of proposals.

The presence of biographical data about the authors was felt as a noticeable improvement. None of the panellists thought that the presence of links to external sites was irrelevant,

though they found it disturbing that the content of the linked pages was not displayed in a new window but substituted to the page on which they had clicked, making it difficult then to realize that they had left the catalogue. They experienced some difficulty in returning to the OpenCat interface after they had followed links to external sites.

Oddly enough, none of the panellists made any comment on the fact that all the editions of a given work are gathered under the uniform title of that work. Perhaps we can regard this absence of particular comments as an indication that this way of organizing bibliographic information was felt by the panellists as so 'natural' that it was not even worth noticing it. Perhaps, in spite of the reduced number of panellists, this can be interpreted as a proof of concept of the FRBR model, showing that searching and identifying a work is perceived by end-users as more important than searching and identifying a specific manifestation, the title of which may reveal nothing as to the various works that can be embodied in it. A data model that would content itself with repeating the manifestation title at the work level would therefore be likely to be deemed inefficient in terms of convenience of the user whenever a manifestation contains more than one work. One of the panellists suggested that a page devoted to a given author should list, not just that author's works, but also works derived by other authors from that author's works – an indication that even complex bibliographic relationships are of interest to library patrons without any training in cataloguing, and that simplistic views about bibliographic information might well miss the point.

5 FOLLOW-UP AND FUTURE DEVELOPMENTS

Although the OpenCat project is not over yet, other local libraries already have expressed their interest in participating in similar projects. The public libraries of Saône-et-Loire (a French department located in Burgundy) wish they could use OpenCat to build a union catalogue at the department level. The idea would be that each individual library in the local network could keep their own internal management system, while sharing a common OPAC based on OpenCat that would provide access to a common dump of semi-FRBRized bibliographic data from the data.bnf.fr Web site. Information relating to the availability status of the holdings of each library would be added to that dump, making it possible for end-users to reserve a document from their mobile phone.

OpenCat can therefore be regarded as a promising project that opens new potentials for library collaboration at the national level. Logilab intends to develop another demonstrator, which will allow libraries to test the display of a selection of records from their catalogue, in order to show that the prototype was not designed to be used by the Fresnes library only but that any library can have it customized in order to meet their needs.

OpenCat was not designed as a tool for creating bibliographic data, but for displaying pre-existing bibliographic data on the Web in a manner that is likely to enhance its usability. It is not a cataloguing tool, but a tool that enables local libraries to customize a common OPAC that provides access to bibliographic data produced at the national level, and holdings data produced at the local level. With OpenCat, BnF made the strategic choice of disseminating its data in a formalism that is compliant with the Semantic Web, even before the production tool that serves to create it was changed.

The advantage of OpenCat is that local libraries can implement it without having to change their ILS. This possibility allows for a period of transition before all libraries of a given country can evolve from a MARC environment to a Semantic Web environment.

6 REFERENCES

Giappiconi, T., (1998). Les ressources bibliographiques de la Bibliothèque nationale de France: la politique bibliographique de la bibliothèque de Fresnes. *Bulletin des bibliothèques de France*, 43(6), 26-33.

Illien, G, (2012) Are you ready to dive in? A case for Open Data in national libraries [PDF]. Paper Presented at the IFLA World Library and Information Congress, Helsinki. Retrieved from: http://conference.ifla.org/past/ifla78/181-illien-en.pdf.

Simon, A., Wenz, R., Michel, V., Di Mascio, A. (2013). Publishing Bibliographic Records on the Web of Data: Opportunities for the BnF (French National Library). In *The Semantic Web: Semantics and Big Data. 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, 563-577. doi:10.1007/978-3-642-38288-8_38. Retrieved from: http://eswc-conferences.org/sites/default/files/papers2013/simon.pdf