Submitted on: 8/5/2014

We grew up together: data.bnf.fr from the BnF and Logilab perspectives

Agnès Simon
Bibliographic and digital information department, Bibliothèque nationale de France, Paris, France
E-mail address: agnes.simon@bnf.fr

Adrien Di Mascio
Data management department, Logilab, Paris, France
E-mail address: adrien.dimascio@logilab.fr

Vincent Michel
Data management department, Logilab, Paris, France
E-mail address: vincent.michel@logilab.fr

Sébastien Peyrard
Bibliographic and digital information department, Bibliothèque nationale de France, Paris, France
E-mail address: sebastien.peyrard@bnf.fr

Abstract:

Three years after the launch of the linked open data site data.bnf.fr, we can report on the experience of the project, from the cross perspective of a public institution, the National library of France (BnF) and a company, Logilab. Starting like a small innovative project, with few data and a small team, data.bnf.fr is now becoming an up-and-running service, progressively integrating all the resources from BnF's catalogues, and with a broad audience.

This paper shares what made data.bnf.fr a success story: librarians and IT services from the BnF and Logilab programmers working together according to agile software development methods; using the free software CubicWeb based on a relational database; relying on the long term cataloguing and diffusion policy of the library.
Yet we are now tackling technical, organizational and strategic issues regarding scalability, dependencies, stability, but also knowledge transfer to newcomers on the project. We are now considering the project in a long term perspective, integrating it to the BnF routines and issues, but also keeping on innovating.

Keywords: data.bnf.fr Cubicweb semantic web organization agile software development matching automation scalability

Introduction

In July 2011, the National library of France launched its linked open data web site *data.bnf.fr*. The goal of the project is to make library data and resources more useful to people and machines on the web. *data.bnf.fr* automatically aggregates BnF resources from the main catalogue, the digital library Gallica, the archives and manuscripts catalogues and other applications, into pages about authors, works and themes, according to the FRBR model. This data is displayed in RDF and JSON, and is available under the Open License that allows any reuse of the data, provided that the BnF source is quoted. Today it is no longer a prototype, but an up-and-running service with users relying on it for daily activities. We can now report on this experience in a "lessons learnt" way: which best practices emerged, which difficulties were met.

The project has been a key for BnF to evolve in a tangible way on semantic web techniques[1], open data[2], and the FRBR[3] model. Yet we won't address the description of *data.bnf.fr*, here, but we will intentionally focus on the back office of the project, on what *makes it happen*: librarians and programmers working together. To develop *data.bnf.fr*, BnF called for tender and adopted the free software CubicWeb, externally developed by a software development company, Logilab that had no experience in libraries, but rather in linked data and web publication. How did the librarians and the IT service from the BnF build *data.bnf.fr* together with Logilab? What were our choices and issues in terms of work organization and of information systems? What did we learn from each other?

The keys to success in building *data.bnf.fr* came from the BnF and Logilab approach of the project and on the work organization. Yet we are facing obstacles and issues in order to make *data.bnf.fr* a "business as usual", as the site is changing and growing. Finally we will try to consider the project in a long term perspective: data.bnf.fr has to be part of the BnF routines, but also has to keep on innovating.

1. Keys to success

The BnF perspective: think user first

Libraries have the mandate to give access to their resources in the easiest way possible. In this regard, using semantic web technologies is not only about exporting raw data, but first about providing services to web users. The goal of *data.bnf.fr* has been to build pages that users can find, browse and quote easily and trustfully on the web. The same URIs can be used to display the HTML page as well as the RDF or JSON underlying data, and the URIs provided in the RDF data can lead to landing pages or to the corresponding raw data, depending on users' needs; finally the HTML data model is similar to the RDF one. We try to comply with web standards providing standardized and stable URIs, and organizing information around entities so that a search engine has a single access points to several pieces of data. We also comply with best practices, like avoiding data duplication by adding "no follow" links. Whenever it was tied to a potential user need, new pages about years (1821) and places (Rouen) were created and the author's pages linked to them to enhance their discoverability. These matches are made with a python open source tool, Nazca, that provides a high-level API for data alignment[4]. Finally schema.org and Open graph protocol data were embedded in the HTML pages, to help search engines and social networks finding the link to digital documents and get the main information from the pages. The aim of this approach is not to artificially inflate our page rank, but to be on the internet user's path and provide relevant answers to his questions. As a library that holds rare or "niche" heritage contents, we bet on Chris Anderson's "long tail" effect: "As they [consumers] wander further from the beaten path, they discover their taste is not as mainstream as they thought (or as they had been led to believe by marketing, a lack of alternatives, and a hit-driven culture)".[5]

Building on the library's existing strengths

This goal being defined, we could rely on existing BnF data to build *data.bnf.fr*. In a sense, BnF has been preparing for the semantic web and *data.bnf.fr* for a long time. Indeed the library has been using and developing structured formats, such as the MARC formats exported in XML formats, EAD in XML. Moreover, it has been creating unique and stable ARK (Archival resource key) identifiers, compatible with HTTP, for bibliographic and authority records and digital documents, and is maintaining a common resolver[6]. It was easy to build stable URIs in the form: http://data.bnf.fr/ark:/12148/cb16192063d.[7]
Furthermore, the library has been developing authority files on authors, works, subjects, or places that can be used as nodal points to organize access to countless resources of various types and structured in different formats: descriptions from different databases are linked to the central authority repository through ARK identifiers. Besides, we already record the nature of these links: for instance, the link between a book and its author or contributor is usually specified by a MARC relator role code, controlled in our repositories. Those pre-existing long-term identifiers, along with a cataloguing policy based on hardcoded record linking is underpinning the strengths and quality of *data.bnf.fr*.

Using a relational database: CubicWeb[8] by Logilab

Then we had to develop the site itself. The library examined different solutions, including triple stores, but finally adopted a relational SQL database, considering the price and the maturity of the technology. It appeared that it was relevant to develop semantic web sites with existing and well-tried technologies for internal management purposes: the library's systems administrators would be familiar to them, which makes it easier to interact with teams who work on related systems (e.g. catalogues) inside the library. From the librarian point of view, it is also an opportunity to keep the production formats and workflows as they are, each format being well-suited to different materials (EAD for archives, MARC format for simple documents, TEI for book binding descriptions). For those reasons, we chose to separate the data production from the dissemination on the web. The SQL database is a hub that merges all the data and serves web pages and raw data in several formats at the same time (RDF/JSON/CSV/HTML). The application uses an internal data model which reflects the needs of the web application and adapts its internal data and data model to standard ontologies. It allowed us to change ontologies when necessary or to duplicate the information with different properties. For instance, we internally define the notion of Person, which is later exposed in two classes that share common properties but also have specific ones: as a skos:Concept describing the record about the person, and as a foaf:Person describing the person itself.

Service-building by team-building: the agile software development experience

This software is developed according to the agile software development: the BnF *product owner* gathers the needs of the library, while the *scrum master* in the IT service deals with technical and contractual issues. The librarian follows the whole workflow of the project, from the development, testing and validation, to the release of the versions. Indeed we need to keep close to the users' requests, and to be able to adapt the site to the web's evolutions. In practice, the separation between these actors is not so strong, as librarians and developers work together to define the roadmap of the product. Developers may propose new functional requirements, while the librarian remains vigilant on technical and performance issues. In practice, BnF and Logilab keep in contact every day through an extranet and an instant messaging tool.

Evolutions are gathered in three week long *iterations*: the requirements are prioritized, and completed or transformed according to feedback from the project team and users. At the end of every second iteration (i.e. every 6 weeks), the evolutions are presented to a group of 20 librarians from different services, and discussed. These feedbacks are crucial to help prioritize the evolutions and the roadmap. For each evolution, we write tests with real data so that librarians can check the evolutions and spot any regressions, especially when a new content type (e.g. music, periodical, performance) is added to *data.bnf.fr*.

At the end of every 5th or 6th iteration (about 3 or 4 months), a new version of *data.bnf.fr* is released. We do not release a new version corresponding to each iteration, but wait to have a consistent set of new features which result in a complete product. For instance, if we want *data.bnf.fr* to handle a new kind of record (e.g. geographical records), we will have different user stories for importing the specific MARC structure, publishing the HTML page, the PDF page, the RDF data, aligning those records to others, etc.

2. Becoming business as usual: the issues

Logilab and BnF are now under a second public contract until 2015. We plan to make *data.bnf.fr* a *business as usual*, and to display all the validated data of the catalogue. As the project is growing, its evolutions and programming is getting less "agile": the code gets more difficult to maintain and to change, releases require more time, and technological constraints are stronger.

Stability and updates

Developing a site step by step and accumulating business rules tend to make the code complex. It needs to be improved regularly in order to improve the performance and stabilize the site. Besides, we have to handle massive, specialized amounts of data with a variable quality level, due to the history of BnF's catalogues. The 17 million bibliographic records and 2 million authorities have been displayed gradually on the web, with now (February 2014) 40% of our catalogues in *data.bnf.fr*. Dealing with a huge amount of data has consequences on performance. On the IT side, a stabilization phase is now scheduled before the release of every new version, especially as new functional requirements are developed while increasing the size of the database. From the user's point of view, loading the pages happens to be problematic, as they have accumulated more content. The audience of the site is also growing, with now 100 000 unique visitors per months, and an increased use of RDF download through content negotiations. There are already technical solutions to display large pages, such as a varnish cache[9] that keeps the page in memory and serves it immediately to the user, but new solutions are needed.

Dependencies and data workflows

Besides we have to import data from several databases, not only from the main catalogue, but also small specialized databases such as *reliures.bnf.fr* or *bp16.bnf.fr*, or data that is not catalogued such as BnF bibliographies or the Web legal deposit information. We are working on automating the workflows of the input of data in *data.bnf.fr*. But it also points out the limitations of the multiplication of sources of data in an institution and paradoxically prompts BnF to coordinate its cataloguing processes.

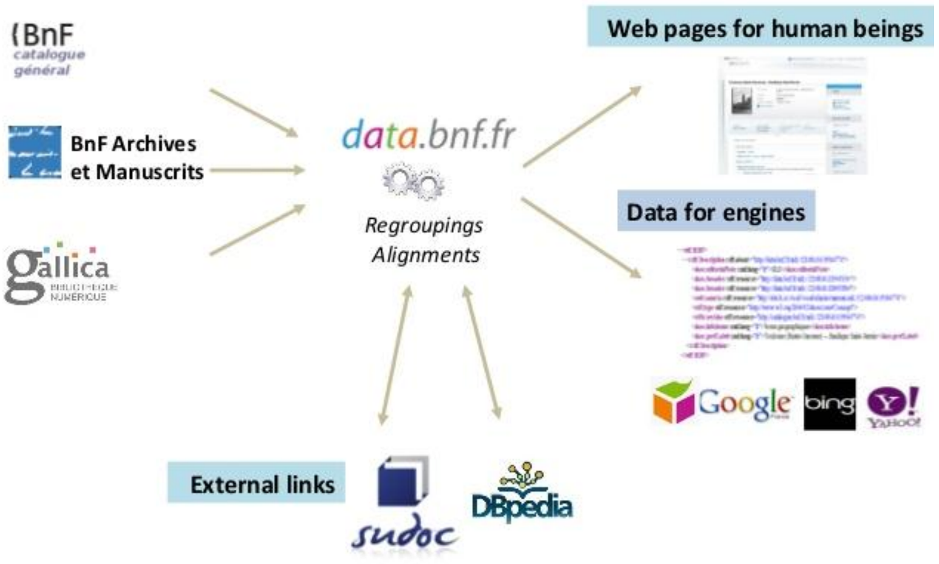


Illustration 1: workflow of data.bnf.fr

data.bnf.fr is also linked to many datasets through VIAF, which plays the role of a hub to DBpedia and other library datasets. We have to be heedful of the changes that may occur within external datasets and to choose broadly used datasets such as Geonames (for example: http://data.bnf.fr/11943054). Dependency problems take new proportions when importing data from the web.

Scaling up: only a computer issue?

To tackle the issue of scalability and stability, the code is refactored and the deployment processes and infrastructure are improved with BnF’s IT services. Yet, this is not only a technical matter: librarians also have to decide on the evolution of the product. For instance we have to prioritize the content that is displayed in order to lighten the main pages, the PDF exports and create secondary pages, for example providing only editions of the Comedie by Dante illustrated by Gustave Doré (http://data.bnf.fr/cross-documents/11900422/11952658/440/page1). An example of simple optimization is also to lighten the RDF content negotiation exports for each page, without losing information. For instance, before the optimization, http://data.bnf.fr/11900422/gustave_dore/rdf.xml provided the whole description of his related documents (extract):

```
<http://data.bnf.fr/ark:/12148/cb30001382m#frbr:Expression>
bnfroles:r440 <http://data.bnf.fr/ark:/12148/cb119004228#foaf:Person>
dc:contributor <http://data.bnf.fr/ark:/12148/cb119004228#foaf:Person>
<http://data.bnf.fr/ark:/12148/cb30001382m>
rdf:type <http://rdvocab.info/uri/schema/FRBRentitiesRDA/Manifestation>
bnf-onto:FRBNF
rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">30001382</bnf-
onto:FRBNF>
dc:description "In-8°"
dc:publisher "1886 Paris Hachette"
rdagroup1elements:placeOfPublication "Paris"
dc:date "1886"
rdagroup1elements:designationOfEdition "Nouv. éd."
rdfs:seeAlso <http://catalogue.bnf.fr/ark:/12148/cb30001382m>
dc:title "Le roi des montagnes"
rdagroup1elements:publishersName "Hachette"
```

Now the URI of the author only provides the links to his related document, for example:

```
<http://data.bnf.fr/ark:/12148/cb30001382m#frbr:Expression>
bnfroles:r440 <http://data.bnf.fr/ark:/12148/cb119004228#foaf:Person>
dc:contributor <http://data.bnf.fr/ark:/12148/cb119004228#foaf:Person> ,
```

the description of the document itself being accessible under the address of the document: http://data.bnf.fr/ark:/12148/cb30001382m.

Thus it is easier to generate the RDF of the author’s page, when the user needs it through content negotiation.

Finally, displaying a large amount of data in huge dumps becomes unsatisfactory for some users that do not have the time or the infrastructure to download and handle a whole dump (100 million triples for 40 % of the catalogue). Solutions are being planned to accommodate those issues: splitting the main dump in smaller dumps about authors, works, manifestations, external links for instance; giving a systematic access to simple exports such as JSON for all *data.bnf.fr*; finally, opening a SPARQL endpoint, so that the users can pick and choose their data in the whole dataset without querying the inner database (on the performance side) or having to handle a dump (on the user side).

3. data.bnf.fr in the long term: routines and innovation

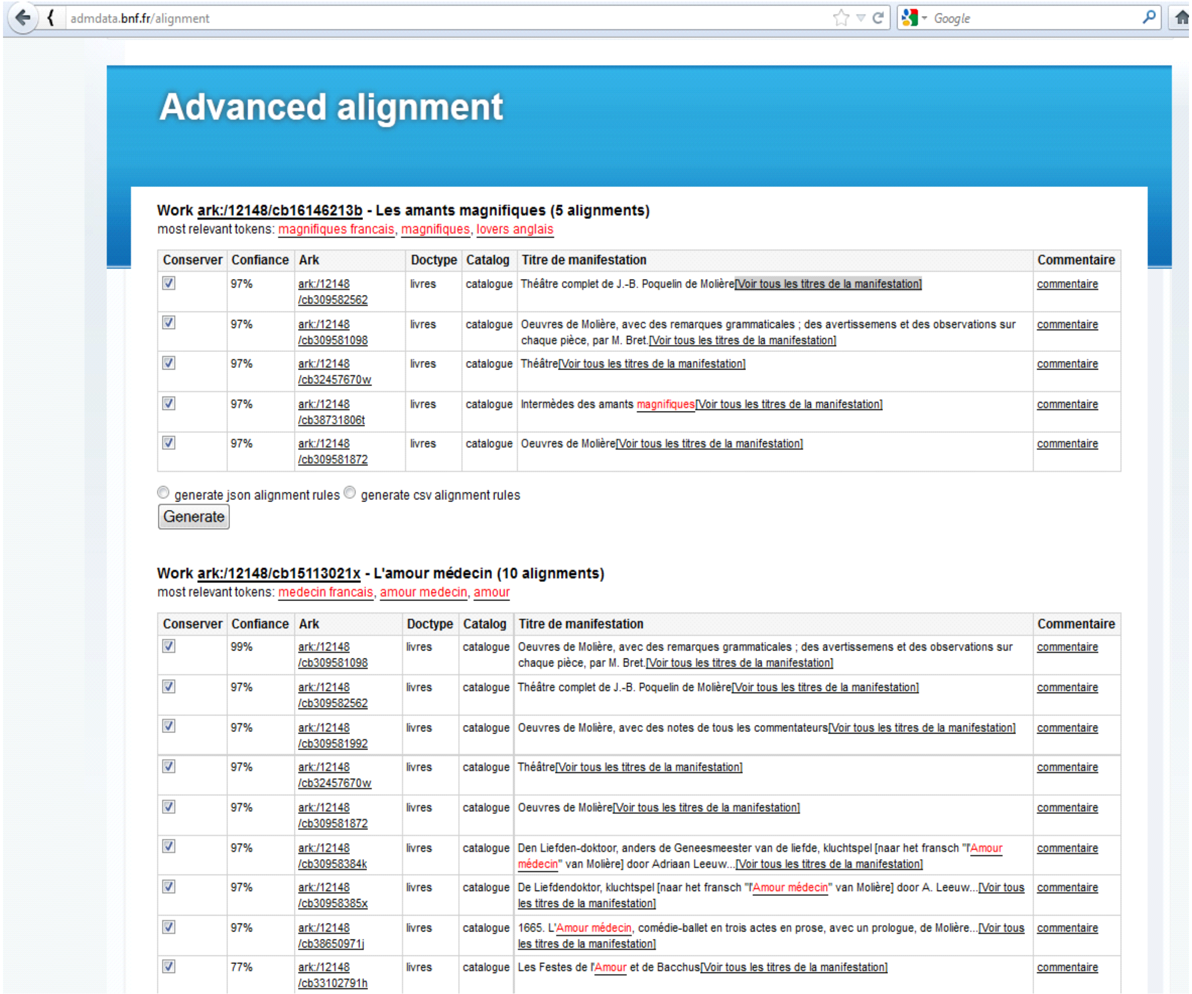
Knowledge transfer

Our first preoccupation is to integrate the application to the BnF IT system and workflow, in a long term perspective. Therefore, knowledge transfer from Logilab to the BnF IT is capital. The CubicWeb software, though it is free and developed in a common language (Python), has to become familiar to our IT services. It also becomes harder for a new programmer to get into the project, as it is now the result of an accumulation of developments since the beginning of the project.

On the librarian side, transferring the project also requires time to be at ease not only with semantic web technologies and technical issues, but also with all the aspects of the project, in particular the strategic ones like convincing people, initiating partnerships. Globally we dedicate a great amount of time to communication and teaching around *data.bnf.fr* and the semantic web inside the library and among the library community.

Consolidating our catalogues with algorithms: *data.bnf.fr* as a FRBRization tool

On the librarian side, *data.bnf.fr* may also help improving our catalogues and move towards a FRBR structure. Logilab developed and used algorithms based on machine learning techniques that facilitate the comparison of massive databases. First we can create links between existing title authority data and the corresponding bibliographic records by rolling back into the catalogue the results of the highly reliable matching processes. Second, for less reliable results, we built an interface targeted at a BnF pool of cataloguers and correctors: for every works of an author, the machine suggests potential links to bibliographic records, which are checked by an operator, exported as CSV and are intended to be processed back into the catalogue to add all the relevant links.



Advanced alignment

Work ark:/12148/cb16146213b - Les amants magnifiques (5 alignments)
most relevant tokens: magnifiques francais, magnifiques, lovers anglais

Conserver	Confiance	Ark	Doctype	Catalog	Titre de manifestation	Commentaire
☑	97%	ark:/12148 /cb309582562	livres	catalogue	Théâtre complet de J.-B. Poquelin de Molière[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb309581098	livres	catalogue	Oeuvres de Molière, avec des remarques grammaticales ; des avertissemens et des observations sur chaque pièce, par M. Bret.[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb32457670w	livres	catalogue	Théâtre[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb38731806t	livres	catalogue	Intermèdes des amants magnifiques[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb309581872	livres	catalogue	Oeuvres de Molière[Voir tous les titres de la manifestation]	commentaire

○ generate json alignment rules ○ generate csv alignment rules
Generate

Work ark:/12148/cb15113021x - L'amour médecin (10 alignments)
most relevant tokens: medecin francais, amour medecin, amour

Conserver	Confiance	Ark	Doctype	Catalog	Titre de manifestation	Commentaire
☑	99%	ark:/12148 /cb309581098	livres	catalogue	Oeuvres de Molière, avec des remarques grammaticales ; des avertissemens et des observations sur chaque pièce, par M. Bret.[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb309582562	livres	catalogue	Théâtre complet de J.-B. Poquelin de Molière[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb309581992	livres	catalogue	Oeuvres de Molière, avec des notes de tous les commentateurs[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb32457670w	livres	catalogue	Théâtre[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb309581872	livres	catalogue	Oeuvres de Molière[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb30958384k	livres	catalogue	Den Liefden-doktoor, anders de Geneesmeester van de liefde, kluchtspel [naar het fransch "l'Amour médecin" van Molière] door Adriaan Leeuw...[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb30958385x	livres	catalogue	De Liefdendoktor, kluchtspel [naar het fransch "l'Amour médecin" van Molière] door A. Leeuw...[Voir tous les titres de la manifestation]	commentaire
☑	97%	ark:/12148 /cb38650971j	livres	catalogue	1665. L'Amour médecin, comédie-ballet en trois actes en prose, avec un prologue, de Molière...[Voir tous les titres de la manifestation]	commentaire
☑	77%	ark:/12148 /cb33102791h	livres	catalogue	Les Festes de l'Amour et de Bacchus[Voir tous les titres de la manifestation]	commentaire

Illustration 2: alignment results for the works of the author Molière

The last step will be to test clustering algorithms to create title authority data when it does not exist. For instance for the work *Le sens du combat*, by Michel Houellebecq, the machine should propose to create a work and link it to the existing manifestation. At the time of this writing these algorithms are being tested on the whole BnF catalogue, to see how well they scale when run on all bnF data. The results of this experiment should tell us if these automatic results should be controlled and validated by a cataloguer through the administration interface or not, according to the confidence level and the record and content type being handled.

Finally those algorithms have also been tested on the archive and manuscripts catalogue to link the EAD finding aids to the authority records of 9000 authors.

These alignments and clustering features, used in *data.bnf.fr* are not only a huge opportunity to improve the catalogue quality; in the light of *data.bnf.fr* sustainability, it is also a way to intertwine more closely very well-established and traditional activities in the library to this new service.

Keeping on innovating

The future of *data.bnf.fr* is not only about maintenance, but also involves carrying on with the innovation and enhancing new uses of its data. We are aware of some uses outside the library community, with applications like Isidore or Abuledu, IFVerso. For the BnF, working with an external company is an opportunity to see its data from the perspective of people who consume it. Logilab proposes new visualizations of the data in a Lab (http://data.bnf.fr/atelier) that are tested and may be integrated to the main pages in the long term. Furthermore the BnF and Logilab proposed a prototype that links the national dataset to local information and to web resources, for public libraries OPACs: OpenCat[10][11]. Finally we use *data.bnf.fr* as a proof-of-concept in order to prompt public institutions to display structured data in open data, in order initiate a virtuous circle of innovation and services building based on new partnerships. That is why we are considering new links to reliable datasets from French institutions, such as geographical and administrative reference datasets like IGN and INSEE, or Cité de la Musique (City of Music) for musical resources.

Conclusion: what did we learn?

On the library side, *data.bnf.fr* triggered many changes. The organization of the project and the way we interact with the developers were quite new to a large public institution like the BnF. *data.bnf.fr* grew within an R&D spirit, with a small team, trying and testing new functional requirements, focusing on its strengths and leaving aside the less successful evolutions. We could work step by step showing tangible evolutions to users and taking into account their feedback. Moreover we learned that we could change our approach of our job as algorithms may help us to correct and improve our data, by combining automatic treatment with human work.

On the company side, Logilab had to learn how to adapt to the needs and processes of a public, huge and very structured institution. Programmers became familiar with library data and issues. Though their core business is not library information systems, Logilab now has assets that meet growing needs for libraries willing to publish their data on the web, to improve access to their resources or to process and match their data.

It appears that the main success of a project like *data.bnf.fr* relies on the commitment of the actors involved in the project according to their specific roles (librarians, the IT service and Logilab), but also on their ability to go beyond these traditional roles. The boundaries of our respective work and roles have moved in this “second machine age”[12].

References

[1] Boulet V. S’appuyer sur la structure des données et les liens : le format MARC comme tremplin pour le Web de données : l’exemple de data.bnf.fr. In : World library and information Congress: 78th IFLA general Conference and Assembly ; 2013 ; Singapore. Available from : http://library.ifla.org/250/1/222-boulet-fr.pdf.

[2] Gildas I. Are you ready to dive in? A case for Open Data in national Libraries. World library and information Congress: 78th IFLA general Conference and Assembly; 2012; Helsinki. Available from: http://conference.ifla.org/sites/default/files/files/papers/wlic2012/181-illien-en.pdf.

[3] Di Mascio A, Michel V, Simon A, Wenz R. Publishing bibliographic records on the Web of data: opportunities for the BnF (French national Library). Lecture Notes in Computer Science. 2013; Volume 7882: pp. 563-577. Available from: http://link.springer.com/chapter/10.1007%2F978-3-642-38288-8_38#page-1

[4] https://www.logilab.org/project/nazca

[5] Anderson C. The long Tail. Wired Magazine. 2004. Available from: http://web.archive.org/web/20041127085645/http://www.wired.com/wired/archive/12.10/tail.html.

[6] http://www.bnf.fr/en/professionals/other_international_identifiers/a.ark_other_identifiers.html

[7] The BnF URI policy based on ARK is specified on http://data.bnf.fr/semanticweb/

[8] http://www.cubicweb.org/

[9] http://www.varnish-cache.org/

[10] Le Boeuf P. Customized OPACs on the Semantic Web: the OpenCat prototype. In: World library and information Congress: 78th IFLA general Conference and Assembly; 2013; Singapore. Available from: http://files.dnb.de/svensson/UILLD2013/UILLD-submission-3-formatted-final.pdf.

[11] OpenCat Prototype. Available from : https://demo.cubicweb.org/library/

[12] Brynjolfsson E, McAfee A. The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. New-York: W. W. Norton & Company; 2014.