Methodological Proposals for Designing Federative Platforms in Cultural Linked Open Data: the example of MoDRef

Antoine Courtin, Jean-Luc Minel

Methodological Proposals for Designing Federative Platforms in Cultural Linked Open Data: the example of MoDRef

Antoine Courtin (University Paris Ouest Nanterre La Défense)
Jean-Luc Minel (MoDyCo, University Paris Ouest Nanterre La Défense, CNRS)

As part of the on-going Labex project `Past in the present_, our proposal aims at highlighting the organizational issues of linked data projects that have to deal with pluri-institutional contexts, among which libraries. First, we will discuss what is at stake. Second, we will present a methodology based on the building of several diagrams which highlight technical, conceptual, and organizational obstacles. We will also address the issues of designing and producing an information system intended to ensure the transmission of scientific skills, the exploitation of major vocabularies, associated to specific vocabularies, by foreign institutions and the harmonizing or building of bridges between heterogeneous descriptions.

The opportunities afforded by digital technologies for digitizing objects, processing digital data and displaying these data and metadata have led patrimonial institutions (libraries, museums, archives) and research units to build `production lines_ and then data repositories. In addition to questions of professional standards, the visibility of these data repositories is an urgent concern within patrimonial institutions at the present time: 2013 was remarkable for the number of initiatives, notably supported by the Ministère de la Culture et de la Communication. In France as in other countries, this mass of projects had led to various proposals being put forward. Some of them focus on building conceptual models, sometimes called ontologies, and can be sharply distinguished from one another by their disciplinary origins.

Our starting assumption is based on the use of languages from the semantic Web, and especially formal languages such as RDF, RDFS and OWL (Berners-Lee et al. 2005), to produce digital tools for data processing, to allow real interoperability between the various repositories created (Hyvönen 2012) and then to form linked data. These tools have not only a documentary ambition but should also give rise to new research perspectives by increasing the sources of information.

At first sight, the crucial concern of patrimonial institutions is the insertion of the data produced by an institution within an ecosystem, beyond disciplinary borders. From an operational point of view, the concept of interoperability is the most widely used (Hyvönen 2012), with a distinction between syntactic interoperability, which deals with data exchange standards and digitized annotations, and semantic interoperability, which concerns the semantics of annotations. We consider, however, that designing a platform should be envisaged through every dimensions of the ecosystem and not only

through interoperability. More specifically, specialized descriptive vocabularies and authority lists are the cornerstone of developments in the patrimonial field. However, these resources depend mainly on the SKOS language, which is in fact used as a standard and has become one of the main languages of the semantic web (Gandon et al. 2012). In January 2014, around 400 specialized vocabularies had been recorded (Vatant et al. 2013). Moreover, the boundaries between conceptual models and vocabularies are not clearly defined: FRBR is designed as a vocabulary (Vatant et al. 2013) although we would argue that it is more a conceptual model. Some ontologies such as the CIDOC-CRM or HADOC (Harmonisation de la production des données culturelles) claim to embrace Cultural Heritage as a whole and to transcend institutional borders (museums, libraries, archives).

In order to tackle this multidimensional complexity, i.e. conceptual, technical and organizational features, we have designed a methodology that is able to both identify obstacles and provide some solutions.

Our presentation will be based on the work conducted since 2012 as part of the project group `Modèles, Référentiels et Culture Numérique_ (ModRef) in the labex `Past in the present_ (http://www.passes-present.eu/). This labex associates 7 research units and 4 patrimonial institutions, including 3 libraries (Bibliothèque nationale de France, Bibliothèque de documentation internationale contemporaine, Bibliothèque Éric-Dampière). The advantage of these libraries for the ModRef project is that they represent various environments, scales, data models and usages and therefore give us a broad vision of the issues involved. Moreover, these issues, which are specific to libraries engaged in linked open data, have to be, in our case, recontextualized with other patrimonial institutions in order to identify the possible conceptual, organizational and technical divergences and convergences. The question is no longer one of considering linked data inside a single institution ´ the library ´ , but of considering it as an exchange institution from the very beginning. By taking this global approach to digital humanities, the ModRef group aims to work on the interoperability of multimodal corpora (sounds, texts, images, videos), which are produced by the partners, and with other data repositories and then to better insert them in the linked data ecosystem. Through it, we wish to show partner institutions the interest of using LOD by highlighting the gains within the labex and inserting them in a long term perspective.

Five descriptive dimensions have to be considered: concept, technologies, organization, licence issues and user experience. Three of them will be developed as examples. These dimensions have nevertheless to be recontextualized with respect to the intended aims of data reuse and enrichment by placing usages at the heart of the thinking. We will focus on the challenges created by the multidimensionality of information within the patrimonial field and by the variety of institutions involved and anticipated end-uses. To handle this multidimensionality, an approach based on the parsing of complex systems appears appropriate (Mitchell 2009).

Concerning the conceptual dimension, we have already mentioned the variety of models resulting from different disciplinary traditions. Those ones concern on the same time the conceptual (or structural) model and the XML syntax. We contend that it would be more advisable to separate the two notions: on the one hand the conceptual model, and on the other the syntactic format with the language used to write the grammar of the model. We have to point out that it is difficult to forecast what the main standard (either normalized or actually used) will be in the five next years. In France, the Bibliothèque nationale de France seems to be tempted by the FRBR, while the Ministère de la Culture promotes the use of the harmonized model created through the HADOC program for producing cultural data, and the Archives Nationales, part of the ICA/EGAD working group, has chosen to create an OWL-ontology. At the European level, Europeana promotes its own EDM model. These are just a few examples among many. As this instability has to be taken into consideration, the aim of our methodological proposals is to ensure the adaptability of the target system. This multiplicity of initiatives, in addition to the question of choice, makes it extremely complex to address the conceptualization of patrimonial information. The pedagogical aspect is also an important element in the organizational dimension and has to be considered in such a project.

The technological dimension is more or less structured, as plenty of digital tools are dedicated to the management and exploitation of digitized data. The functionalities of these tools are broad, ranging from cataloguing to the management of enquiries, the integration of data curation and the maintenance of thesauri or authority lists. By considering only the technical part, these tools can be fully proprietary. This implies in general a dependence on ad hoc or totally open formats and on database management systems with a more or less common format. Often, these two possibilities coexist side by side within the same institution. We argue that the main criterion has to be the exchange model or rather, the possibilities of displaying and then reusing the data, that are offered by existing systems. This exchange model includes four levels. The first level concerns simply the possibility of exporting or dumping data to a more or less common format, generally XML. The second one deals with the exposition of data through a protocol such as OAI-PMH. The third level offers a single public and documented API (application programming interface), while the last one gives access to a SPARQL-endpoint according to W3C recommendations and guarantees insertion within the LOD system.

Our work at ModRef is a good illustration of these various degrees of involvement for data exposition. This technological choice arises not from a technical challenge but from the conjunction of multiple organizational and features which join altogether, as it seems, in order to overstep the data appropriation and also to expose them.

The organizational dimension concerns both the description of the power of each organization (patrimonial institutions, research units and supervisory authorities) and the description of the participants (curators, researchers, users). This dimension is the hardest one to perceive since many

factors that are difficult to categorize come into play, as they are part of a dynamic process (Reix et al. 2011).

In order to find a solution which could be adapted to the needs and constraints of the partner institutions, we will produce several diagrams. Each of them will represent the different dimensions mentioned, to make the conjunction or the dispersion of the projects easy to visualize.

In order to build these diagrams, we have designed a workflow. In the first step, information was collected in the usual way by asking the members involved to fill in analytical grids during interviews. In the second step, the grids as a whole were submitted to all the partners for validation, thus enabling them to correct or complete some features of their own project thanks this comprehensive overview of all the projects. While this two-step approach reduces the process of collecting and consolidating information but it takes part of the organizational learning. The diagrams are created semi-automatically. For obvious reasons of flexibility and reuse, a processing line has been implemented, so that some actions can be automatized. All the data are entered in a spreadsheet and then classified and quantified. A JSON file is then generated to make data manipulation easier thanks to the D4.js javascript library which also makes it possible to create SVG exports that are easily handleable with Inkscape. Figures 1 and 2 show two examples of the diagrams produced.

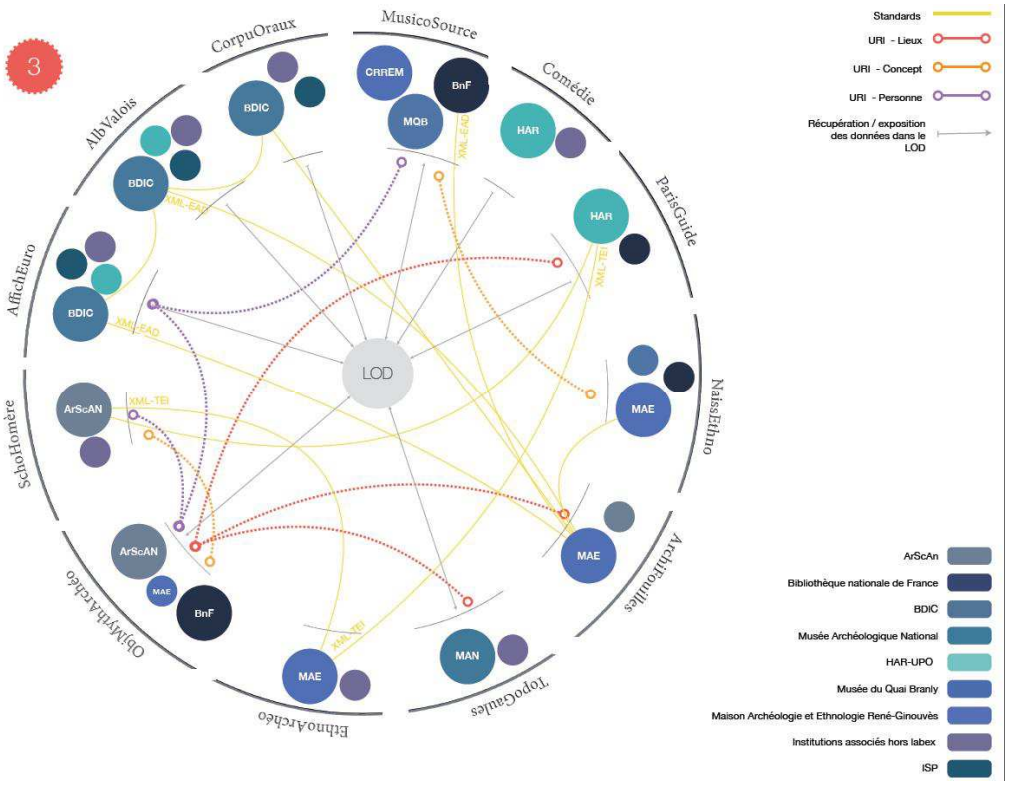

Figure 1: Labex ecosystem

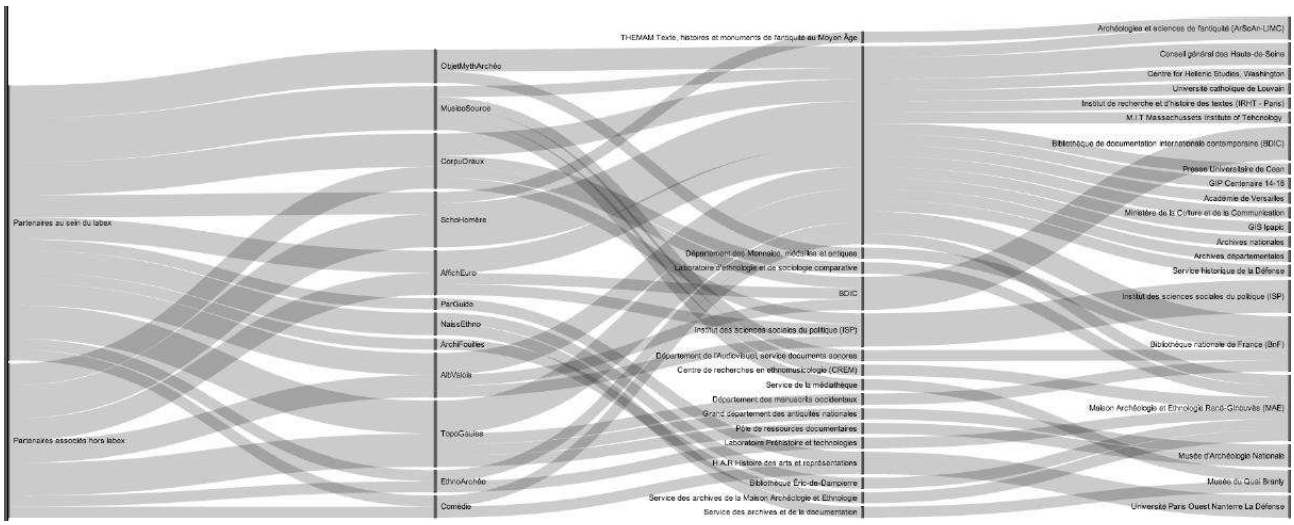

Figure 2 : Organizational complexity

This methodology makes it possible to identify the possible obstacles and to integrate the different participants thanks to a pedagogical approach (training sessions, frequent workshops, heuristic diagrams), but above all to make recommendations to each of our partners. This is the basis of the production of a proof of concept which will start in early April.

Bibliography:

[BERNERS-LEE et al., 2005] BERNERS-LEE, Tim, DIETER, Fensel, HENDLR, James, & LIEBERMAN, Henry, `Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential_, Édition, Wolfgang Wahlster The MIT Press, 2005.

[HYVÖNEN 2012] HYVÖNEN, Eero, Publishing and Using Cultural Heritage Linked Data on the Semantic Web , Morgan & Claypool Publishers, 2012.

[GANDON et al., 2012] GANDON, Fabien, FARON-ZUCKETR, Catherine & CORBY, Olivier, O. Le Web sémantique : comment lier les données et les schémas sur le web ?, Dunod, 2012.

[MITCHELL 2009] MITCHELL, Melanie, Complexity: A Guided Tour, Oxford: Oxford University, 2009.

[REIX 2011] REIX, Robert, FALLERY, Bernard, KALIKA Michael, & ROWE Frantz, Systèmes d'information et management des organisations, Paris, Vuibert, 2011

[VATANT et al.,] VATANT Bernard, VANDEBBUSSCHE, Pierre-Yves (2013), `Linked Open Vocabularies _, http://lov.okfn.org/data (consulted on 4th March 2014)