# Library Metadata on the Web: The Example of Data.Bnf.Fr

Raphaëlle Lapôtre

# Library metadata on the web: the example of data.bnf.fr

Raphaëlle Lapôtre, National Library of France

In library catalogs as much as on the web, metadata acts as a transparent language through which concrete objects are looked for, found and obtained. Past decades have seen web browsing evolving so as to, by reducing query time to an imperceptible amount, make web surfers forget that they are looking for things online: through this, online querying has become as trivial and invisible as would be the daily exchange of money against goods when shopping.

In fact, metadata sets and currencies share in the contemporary world a common nature, or at least common principles: both are measuring the value of things in regard to people's need (of which search engine queries can be seen as a manifestation), and both can be seen as temporary substitutes to those needed things before their obtention.

Such a materiality of metadata is made ever more visible through Open Data policies aiming at pushing them outside of professional communities by transforming them into tradable commodities: records disseminated on the data.bnf.fr website are the product of a thorough cataloging process of which the exhaustivity and continuity is legally grounded by the Legal Deposit mission of the national Library of France. Similar datasets comparisons allowed by the adoption of Linked Open Data standards has been revealing the variety of policy guidances among issuing institutions, demonstrating as such the very archival nature of cultural metadata. It can be said accordingly that catalog records, by recording the activity of their producers as much as cultural collections, constitute historical documents in the same manner as the documentary objects they're likely to describe in a library environment.

It just so happens that Linked Open Data is precisely a way of regarding metadata elements as things to be sorted and named as well as the items being sorted and named through them. The different kinds of reuse of catalog records expressed through such open standards somehow confirm both the materiality of data and ways of functioning of those standards, adding a "meta" prefix to the "meta" of metadata.

The following article will thus examine both the transparency and materiality of bibliographical data firstly by demonstrating its monetary aspect, secondly by showing its archival nature, and thirdly by explaining how adoption of Linked Open Data standards contributes to the reification of data itself. Its argumentation is for its bigger part drawn from the data.bnf.fr experiment, which has been launched in 2011. This website is a project of the national Library of France, aiming at disseminating data from the BnF various catalogs and applications, while constituting a single point of access for users from the web to collections descriptions scattered across many search tools. As this website principles are based on both online visibility and linked open data dissemination, this article will try to demonstrate how those two principles are going hand in hand.

## Metadata, a currency in the attention economy

The fact that when used on the web, metadata sets can be regarded as a currency in the attention economy is especially noticeable when observing the functions of information displayed on a Google results page: just as a traditional currency is supposed to be a substitute for real things while expressing its value among all the tradable commodities, metadata plays on the web a similar role. We can see as a matter of fact that a result page is composed on one hand of short descriptions which are used as temporary substitutes for things that they describe, on the other hand of a strict sequencing in which description ranks are supposed to express the estimated value on the web of the object represented.

As for attention economy, it could roughly be described as the paradox of a limited human sight having to process an almost limitless amount of available information on the web. To that extent, it should be noted that two distinct but complementary ways of dealing with this paradox seem to be predominating on the web: one is handling attentional value by managing human subjectivity, while the other one tries to do the same by synthesizing the huge amount of data appearing on the web. In the first case, this would be the Google-Amazon approach to attention economy, in the second, this would rather be the one of traditional library catalogs.

Our purpose here is to describe with more details those two distinct strategies when used in such metadata functions on the web as representing objects and expressing their values within the set constituted by all of the other items availables for searching.

### Expressing the value of things

As seen before, a way of expressing the value of things is to measure them through comparison in order to sequentially arrange them, so that their relative rank or position in an almost infinite list is indicating its value in regards to the predetermined retrieval criteria.

As for search engines, human subjectivity is the comparison measure used to produce such a serial arrangement of web pages: indeed, the Google algorithm PageRank "works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.[1]" That way of ordering results of search is itself directly inspired from the bibliometric Science Citation Index whose purpose is to classify science articles according to the number of citations they receive, instead of by using the semantic importance of their content as a measure.[2]

In a library environment, such a sequential ordering of documents, visible on its catalog, would have been obtained through the comparison of univoque resource variables such as name of author, title, date of publication, subject, etc. Thus when browsing the shelves of a library, one tries to localize an item with the help of subject classification and alphabetical order, the latter two being a sequential arrangement very much equivalent to a search engine ordering of results, except that it is based on object description rather than relationships between authors of documents.

### Representing things on the web

---

[1] "PageRank". *Wikipedia*. Accessed March 18 2017. https://en.wikipedia.org/w/index.php?title=PageRank&oldid=770098009.

[2] (Cardon 2013, 69-70) Cardon, Dominique. 2013. "Dans l'esprit du PageRank." *Réseaux* 177:63-95.

It is one thing to organize items in order to help patrons retrieve them when they know exactly what they are looking for, it is another one to bring them closer to how those users would *designate* them in their own initial representation of informational need. In a Foucauldian perspective, the above-mentioned term of "designation" could be used in our industrial social computing era to describe two features: first the way informational resources can be clusterized into groups, then the way those same groups of resources can in turn be represented by named entities[3] in a knowledge graph standing for those clusters. The overall purpose of such a substitution is to make online search more intuitive to users, i.e. more acute to mainstream representation of things. In such features again, a subjective-utilitarian approach coexists with a more object-oriented one.

In an article entitled "Algorithmic subjectivity and the need to be informed", the author Neal Thomas describes how such a clustering algorithm as *k*NN is organizing information and knowledge on the web:

> The algorithm and others like it are at work mainly in collaborative filtering services: Netflix's movie suggestion service, Amazon's Recommended for You, and news aggregator sites like Digg and Reddit. *k*NN is a good example of how, through the modern environment of industrial social computing, informational need can be algorithmically structured as continuously intentional, and at least nominally intersubjective: perpetually reorganizing a "neighborhood" of records for present users according to paths laid down by prior ones.[4]

Thus, Google and other web giants are constituting clusters of web pages according to past consultations of users. Such a process provides for the suggesting of related pages when terms are entered into the search bar, allowing the web service not only to give results matching the request but also hints towards next searches[5]: this is how the Google Knowledge Graph is able to suggest related searches in an upper-right infobox when a named entity such as *Les Misérables* is entered in the search engine.

In a similar objective but with opposite methods, the library community is also striving to group resources under named entities. Whereas Google is constituting clusters according to user preferences, allowing for *Gavroche* to be found in the same usage group as Léon Tolstoï's *War and Peace*, the librarian approach consists in building work families complying with bibliographical models such as the Functional Requirements for Bibliographic Records (FRBR - LRM). Here, algorithms similar to the OCLC Work-Set algorithm[6] are not assorting resources depending on user circulation on the web, but by a comparison of character strings made of titles and author names. This is, by the way, precisely how the data.bnf.fr website organizes metadata, namely that each named entity is dedicated a web page on which related documents have been gathered: thus,

---

[3] "In information extraction, a named entity is a real-world object, such as persons, locations, organizations, products, etc., that can be denoted with a proper name." Named entity, 2016. *Wikipedia* . Accessed March 18 2017. https://en.wikipedia.org/w/index.php?title=Named_entity&oldid=747718935.

[4] (Thomas 2012, 9-10) Thomas, Neal. 2012. "Algorithmic subjectivity and the need to be in-formed". In *TEM 2012: Proceedings of the Technology & Emerging Media Track – Annual Conference of the Canadian Communication Association (Waterloo, May 30 - June 1, 2012)*, edited by Gauillaume Latzko-Toth and Florence Millerand. http://www.tem.fl.ulaval.ca/www/wpcontent/PDF/Waterloo_2012/THOMAS-TEM2012.pdf

[5] "I actually think most people don't want Google to answer their questions […] they want Google to tell them what they should be doing next (Eric Schmidt, quoted in Holman, 2010)." (Thomas 2012, 2).

[6] (Hickey and Watts 2009) Hickey, Thomas B. and Jenny Toves. 2009. "FRBR Work-Set Algorithm. Version 2.0." Dublin, OH: OCLC Online Computer Library Center, Inc. (Research division). Published online at: http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf.

the data.bnf.fr page consecrated to *Les Misérables* presents a set of records describing the various editions of that work, instead of links towards usage-similar resources.

Finally, we can observe that whatever the approach adopted for clusterization, named entities literally function as coins or substitutes which, when entered by users in search engines, are exchanged against a set of described objects. What is more, this monetary aspect of metadata is particularly emphasized on the data.bnf.fr website insofar as the last one is making entities and record descriptions available for recovery and reuse. Indeed, this fact contributes to constitute metadata not only as a transparent virtual key towards things, but also as an autonomous object in itself which can be traded similarly to any other device, a feature that it shares with currencies.

## Metadata, a currency in a market of currencies

As for any kind of currency, the quality of metadata in regards to its ability to both correctly measure and stand for objects that it describes, is best measured according to trust, weight and authenticity criteria. The opening of data is precisely what makes metadata sets evaluation possible, as the diversity of their sources allows for comparisons between them.

### Institutional weight

Authority and bibliographic records which are made available on the data.bnf.fr website are for their most part the product of the cataloging activity of the national Library of France. This fact has an impact on the way potential reusers perceive those metadata sets when weighed against other comparable sets. Indeed, as the legal authority of a state guarantees the validity of the currency issued by it, the BnF's legally grounded existence is a guarantee of sustainability over the data sets issued by this organization. It also constitutes a guarantee of comprehensive coverage of all French publications, insofar as the BnF fulfills a role of Legal Deposit for the entire French territory, a role which necessarily implies the description of the entire range of deposited materials.

Moreover, it is to be noted that the description of those collections is traditionally carried out by professionals theoretically with good knowledge both of the kinds of materials they catalog as well as of the description standards to which they must comply while doing so. Such professional cataloging is a sign of quality over metadata production that other data providers such as DBpedia are not necessarily able to provide, whether because of the use of automation process or of a volunteer workforce.

### Validation process

Another appreciated feature of data sets issued by the BnF and disseminated through its data.bnf.fr website is the fact that they have undergone a validation process. In fact, statements that can be found in BnF authority records are most often attached to references of their sources, which themselves come from materials or reference manuals used by catalogers: those references are fully visible on the corresponding entity page in data.bnf.fr, in the "Sources and References" section. Such a sourcing requirement is especially welcomed by Wikidata reusers: indeed, the collaboratively edited knowledge base acts as a metadata hub allowing for comparisons between sets based on identifiers and sources, systematically supporting credit mentions in its own RDF-based data structure.

Although undermining the aforementioned principle of exhaustive coverage, it is yet in the policy of the data.bnf.fr website to avoid dissemination of metadata which is not regarded as validated by the institution: most often, metadata sets lacking a valid status are bibliographic records for which the described material has not yet been deposited, or authority records for which the amount of data elements wouldn't allow for distinction from potential or existing duplicates. If it is true that as of now such a rigorous checking process creates an important differential between the metadata production and its dissemination, this process is nevertheless as critical to data sharing as would be the verification of an alloy in currencies prior to their issuance.

### Benchmarking

A central feature of the data.bnf.fr project is its compliance to semantic web standards. In fact, Linked Open Data offers two advantages: first of all, it provides for the possibility of univoque identification of bibliographic and authority records, and secondly, it makes the unambiguous expression of relationships between datasets possible. It is on this basis that a relationship of equivalence between the BnF identifier of Jean Jaurès and the equivalent entity for the Biblioteca nacional de España (BNE) can be declared, or between a location identified by the BnF and that same location for the French National Geographic Institute (IGN).

Thus, similarly to coins that would be weighed against their equivalents in a foreign country, identifiers and links between datasets allows for a comparison between metadata sets related to common objects: this type of comparison reveals the richness or poverty of data formats, accuracy or inaccuracy of descriptions, or even simply differences in adopted cataloging policies. An example of it can be found in the Virtual Authority File (VIAF) project, which operates as an identifier hub and promotes the linking of identical authority records across the world.

Such reconciliation principle could be used by researchers to study the history of cultural institutions through their data policies, and in order to get a better understanding of the nature of metadata as an anthropological device: that kind of reuse, oriented towards the issuing institution rather than the described object, is exemplified in Loukissas's interactive essay untitled "Life and Death of data"[7], where the very historical nature of metadata is highlighted.[8] Such Digital Humanities projects tend to see metadata sets as material objects, "artifacts" that can speak about the framework of their emergence. It is yet to be seen how such a materiality can be organized and represented through semantic web standards.

## Reuses of Linked Open Metadata: an attempted typology

The first part of this article aimed to describe two main functions of metadata on the web, namely expressing value of real-world things by ordering them on a visual scale on one hand, and being online substitutes for them through their clusterization and designation under named entities on a second hand.

---

[7] (Loukissas 2015) Loukissas, Yanni. 2015. "The Life and Death of Data." Accessed April 10, 2017. http://lifeanddeathofdata.org/.

[8] "The long and storied traditions of record keeping at libraries, museums and arboreta make these institutions opportune sites for studying entangled social and technical changes in practices with data." Ibid.

Interestingly enough, those two functions of measuring and designating can be found again in the very principles of Linked Open Data, i.e. "a method of publishing structured data so that it can be interlinked and become more useful through semantic queries."[9] Instead of measuring and designating objects of the real world, semantic web principles tries to do such things with metadata elements.

## A scale of openness

Undoubtedly, the assertion that open standards can be seen as a way of ordering data elements through measurement would seem rather confusing at first. However, it begins to make sense when we think metadata components being measured not according to attention value but according to a common and unambiguous quality of data element, i.e., its intrinsic openness level. In fact, this very scale of openness was literally defined by Tim Berners Lee in a 2010 W3C web page titled "Is your Linked Open Data 5 stars ?"[10], which describes five essential steps towards open-data:

| | |
|---|---|
| ★ | Available on the web (whatever format) *but with an open licence, to be Open Data* |
| ★★ | Available as machine-readable structured data (e.g. excel instead of image scan of a table) |
| ★★★ | as (2) plus non-proprietary format (e.g. CSV instead of excel) |
| ★★★★ | All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff |
| ★★★★★ | All the above, plus: Link your data to other people's data to provide context |

Figure 1. The scale of data openness, according to Tim Berners-Lee.

Thus, according to this scale, one of the first steps is the use of a data structure to express information, a table being an example of such a structure. In fact, the very principle of a data structure is to organize information following an arbitrary order in which the order number will be decisive to easily retrieve the information. Here too can be observed the same pattern of expression of value through sequential ordering already mentioned in relation to search engine results, except that what is ordered is data elements instead of things described by metadata:

[9] "It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried." "Linked Data." *Wikipedia*, March 24, 2017. https://en.wikipedia.org/w/index.php?title=Linked_data&oldid=771924792.

[10] (Berners-Lee, 2010) Berners-Lee, Tim. 2010. "Linked Data - Design Issues." Accessed April 11, 2017. https://www.w3.org/DesignIssues/LinkedData.html.

position within such an order can be crucial for expressing value of what is ordered, whether regarding human desire (as for things) or openness (as for data). As to the latter, it turns out that RDF serialization is precisely the data structure promoted by W3C, which arrange data elements in ternary structures (subject, predicate, object) in which the central predicate position is occupied by a property (kind of data) necessarily common to several other triples. Indeed, it is another W3C recommendation that the predicate should express a property belonging to a web ontology, which is a set of commonly defined entity types (classes, such as a place or a person) and attributes (properties, such as a name or a geolocation). Accordingly, the property declared as a predicate of a triple necessarily determines the class of the subject and object of the same triple, hence their level of generality in the hierarchy preliminary determined by its ontology: the author Victor Hugo is an instance of the predefined class "People", which itself is a subclass belonging to the general class "Agent". Thus, it can be said that in a Linked Open Data environment, central position in a triple provides clues onto the level of generality, hence openness of the entire triple: the more general a class and its properties, the bigger the necessity of a commonly and collaboratively defined standard to express it.

## URIs and discovery-oriented reuses

As for Uniform Resources Identifiers (URIs), which are among the last elements in Tim Berners Lee's scale of openness, they function in Linked Open Data as keys potentially substitutable for metadata sets linked to them through properties. In fact, it is in the semantic web philosophy to use identifiers as unambiguous "names" for things. Nevertheless, as in other existing LOD projects, the data.bnf.fr linked data model establish a distinction between the real-world thing and its catalog record, so that the first belongs to a class "People" while the latter is labelled as a "Concept" and use a slightly different identifier. Consequently, it can be said that URIs less are names for things than names for distinct sets of metadata attached to a thing. Moreover, similarly to the previously mentioned function of named entities as virtual triggers for sets of real-world objects looked for on the internet, the "HTTP" prefix on URIs bestows the same role to them, providing the user a web page displaying RDF properties and attributes attached to the URIs.

Such a comparison between named entities in web of documents and URIs in semantic web is even more meaningful when we look at a typology of reuses of metadata sets disseminated on the data.bnf.fr website. Up to now, empirical observations have suggested that reusers aiming at building data-displaying interfaces would tend to be more interested in the entity-relationship modeling of metadata sets, while reusers aiming at producing statistical studies relying on data would be more attracted by data set's content itself. In practice, this would mean that discovery-oriented reusing would extract more links of an URI towards other URIS than links of an URI towards descriptive informations, being only interested in a superficial amount of properties describing entities, when research-oriented reusing would be more likely to favour an exhaustive collection of such properties in order to conduct thorough statistical analysis on it.

As for discovery-oriented reutilization, an example of LOD project focusing more on the data networked structure than its content is the Aemoo search interface, which aims at encouraging knowledge exploration by providing a node-link visualization of entities described with a

summary of relationships attached to it[11]. In that particular case, it can be clearly seen that the interface is displaying only a sufficient enough amount of properties for feeding further exploration of related entities, rather than providing a complete view of all links attached to the first entered entity. In contrast, such a research-oriented reusing as the exploratory analysis conducted by Glorieux in a blog post untitled "Data.bnf.fr, les documents"[12] shows a very particular attention being paid to "dead-end" properties of entities such as title, date, language, place of publication or even contributory role on a document, rather than to properties linking the entity towards nodes.

Provided that the very monetary aspect of metadata as part of the web and the semantic web environment has hopefully been demonstrated in the above sections, one could conclude that the node-link structure of interconnected entities would tend to be valued by reusers viewing metadata as light circulatory device allowing for web explorations, while the complete serial arrangement of properties through rdf serialization would be more sought after by researchers looking for heavy gold through data mining.

## Conclusion: from cataloguing things to cataloguing data, among other things…

Through this article, we have explored a combination of binary levels of representation of cultural or non-cultural items.

Firstly, whether metadata is regarded as a transparent device on the web or as documents in itself, functions that are assigned to it are always twofold. As a first step, real-word objects or metadata sets themselves are mapped into a meaningful sequential arrangement, and as a second step, another mapping happens through the respective grouping of those same data sets under named entities on one hand, and URIs on a second hand.

Secondly, with this first duality of web representation comes another one, since while metadata is standing for things on the web and in catalogs, their expression under Linked Open Data standards is promoting for it to have identifiers standing for itself. This is precisely the reason why, when looking at one of these widespread node-link visualizations of RDF data, one can see the representation not only of the real-world items such as an author, his works and collaborators, but also of the very model of data used to represent those items, the node-link diagram reproducing in itself the entity-relationship modelisation of the metadata.

---

[11] "This paper presents a novel approach to Linked Data exploration that uses Encyclopedic Knowledge Patterns (EKPs) as relevance criteria for selecting, organising, and visualising knowledge. EKP are discovered by mining the linking structure of Wikipedia and evaluated by means of a user-based study, which shows that they are cognitively sound as models for building entity summarizations. A tool named Aemoo is implemented which supports EKP-driven knowledge exploration and integrates data coming from heterogeneous resources, namely static and dynamic knowledge as well as text and Linked Data." (Nuzzolese, Presutti, Gangemi, Peroni and Ciancarini, 2016) Nuzzolese, Andrea, Giovanni, Presutti, Valentina, Gangemi, Aldo, Peroni, Silvio and Ciancarini, Paolo. 2016. "Aemoo: Linked Data Exploration Based on Knowledge Patterns". *Semantic Web* 0:1-28. Accessed April 6, 2017. http://www.semantic-web-journal.net/content/aemoo-linked-data-exploration-based-knowledge-patterns-1

[12](Glorieux, 2016). Glorieux, Frédéric. 2016. "Data.bnf.fr, Les Documents." *J'attends Des Résultats*. Accessed April 6, 2017. https://resultats.hypotheses.org/795.

Such an accumulation and intertwining of levels of representation of reality can have the pervert effect of hiding the primary and main goal of cataloging materials behind the complexity of an opening objective which all in all, turns out to be the cataloging of data itself. In this regard, the fact that the national Library of France chose to maintain MARC as a production format while allowing RDF data conversion only for dissemination purpose is not a trivial one: it seems that an institutional distinction between object-focused cataloging and metadata-focused sharing would keep description of library collections safe from taking the word for the thing it stands for.

## References


- (Berners-Lee, 2010) Berners-Lee, Tim. 2010. "Linked Data - Design Issues." Accessed April 11, 2017. https://www.w3.org/DesignIssues/LinkedData.html.
- (Cardon 2013, 69-70) Cardon, Dominique. 2013. "Dans l'esprit du PageRank." *Réseaux* 177:63-95.
- (Glorieux, 2016). Glorieux, Frédéric. 2016. "Data.bnf.fr, Les Documents." *J'attends Des Résultats*. Accessed April 6, 2017. https://resultats.hypotheses.org/795.
- (Hickey and Watts 2009) Hickey, Thomas B. and Jenny Toves. 2009. "FRBR Work-Set Algorithm. Version 2.0." Dublin, OH: OCLC Online Computer Library Center, Inc. (Research division). Published online at: http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf.
- "Linked Data." *Wikipedia*, March 24, 2017. https://en.wikipedia.org/w/index.php?title=Linked_data&oldid=771924792.
- (Loukissas 2015) Loukissas, Yanni. 2015. "The Life and Death of Data." Accessed April 10, 2017. http://lifeanddeathofdata.org/.
- Named entity, 2016. *Wikipedia* . Accessed March 18 2017. https://en.wikipedia.org/w/index.php?title=Named_entity&oldid=747718935.
- (Nuzzolese, Presutti, Gangemi, Peroni and Ciancarini, 2016) Nuzzolese, Andrea, Giovanni, Presutti, Valentina, Gangemi, Aldo, Peroni, Silvio and Ciancarini, Paolo. 2016. "Aemoo: Linked Data Exploration Based on Knowledge Patterns". *Semantic Web* 0:1-28. Accessed April 6, 2017. http://www.semantic-web-journal.net/content/aemoo-linked-data-exploration-based-knowledge-patterns-1
- "PageRank". *Wikipedia*. Accessed March 18 2017. https://en.wikipedia.org/w/index.php?title=PageRank&oldid=770098009.
- (Thomas 2012, 9-10) Thomas, Neal. 2012. "Algorithmic subjectivity and the need to be in-formed". In *TEM 2012: Proceedings of the Technology & Emerging Media Track – Annual Conference of the Canadian Communication Association (Waterloo, May 30 - June 1, 2012)*, edited by Gauillaume Latzko-Toth and Florence Millerand. http://www.tem.fl.ulaval.ca/www/wpcontent/PDF/Waterloo_2012/THOMAS-TEM2012.pdf