Date submitted: 3/08/2010

Porting library vocabularies to the Semantic Web, and back
A win-win round trip

Bernard VATANT
Senior Consultant
Mondeca
Paris, France
http://www.mondeca.com
E-mail: bernard.vatant@mondeca.com

Meeting: **149. Information Technology, Cataloguing, Classification and Indexing with Knowledge Management**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract :

The role of vocabularies is critical in the long overdue synergy between the Web and Library heritage. The Semantic Web should leverage existing vocabularies instead of reinventing them, but the specific features of library vocabularies make them more or less portable to the Semantic Web. Based on preliminary results in the framework of the TELplus project, we suggest guidelines for needed evolutions in order to make vocabularies usable and efficient in the Semantic Web realm, assess choices made so far by large libraries to publish vocabularies conformant to standards and good practices, and review how Semantic Web tools can help managing those vocabularies.

Introduction

Despite the common metaphor of the Web as the Ultimate Global Library, the Web search technologies and the library classification systems had long been considered as either competitors in the knowledge organization market, or addressing independent users niches. The short history of the Web (less than twenty years) compared to the centuries of library experience, the quick raise and fall of many Web ventures, and the crazy life cycle of information technologies have been looked at by librarians with a mixture of contempt and fascination. Nevertheless the notion has slowly but steadily emerged that cooperation and synergy between the "old" library systems and the "new" Web technologies was desirable, and likely to eventually bring about many benefits to both parties. Beyond spectacular technical achievements in data federation and cataloguing such as WorldCat [1], the Semantic Web [2] pile of languages and technologies is bringing about a growing wealth of new opportunities for valorisation of libraries legacy, new perspectives on classification methods polished by library science, as well as new legacy management tools.

Among various aspects of this semantic convergence, we will focus on the central role of controlled vocabularies. We first recall how vocabularies fit in the general evolution of the Web and information systems, then explain why and how the Semantic Web should leverage existing vocabularies instead of reinventing them, then try to assess which features of such vocabularies make them more or less portable to Semantic Web use, and finally suggest guidelines to make them more efficient in this new realm.

Practical examples, issues and lessons learned are based on an experience in the framework of the TELplus [3] project, where the French National Library (BNF) called for a consortium of Semantic Web companies to benchmark a chain of semantic annotation and search tools, integrating large libraries vocabularies such as RAMEAU and LCSH, as well as mappings between those vocabularies.

The other way round, we point out benefits of using Semantic Web tools and software to manage vocabularies, based on our experience with EUROVOC [5] and other vocabularies managed and published by the Publications Office of the European Union.

We end up by a review of current efforts towards global organization of this convergence, the leading role of the Library of Congress Authorities and Vocabularies [4], convergence of various projects in the European Community framework, and the W3C Library Linked Data Incubator Group initiative.

Vocabularies on the Web, a love and hate story

Controlled and structured vocabularies such as classifications and thesauri had been developed by librarians over centuries to support *human indexing* and search, whereas search on the Web has been more and more supported by automatic engines using *statistical algorithms* based on smart full-text search. Search engines do not need structured vocabularies to perform their task, and the tenants of the search technologies have long supported the idea that automatic indexing and search could reach a high level of performance based on smarter and smarter algorithms, and have been steadily declared as more or less obsolete the traditional library classification tools and methods, at least at the scale of the Web. On the other side, librarians and specialists of knowledge organization stuck to the opinion that search engines could not deal properly with ambiguity of terms, and that intelligent search should leverage structured knowledge and human added value embedded in structured vocabularies.

The raise and raise of search engines

The debate was quite open during the first decade of the Web, when the level of performance of search engines was still poor [6]. But the impressive raise and success of Google during the second decade of the Web seemed to have brought a killer argument to the automatic indexing and search camp. Based on this success, some have pushed the argument as far as to declare the obsolescence of controlled vocabularies at the benefit of search technologies. Of course, obsolescence of vocabularies would mean obsolescence of metadata using them, as well as classification altogether, and at the end of the day, the death of library science as it were.

Humans don’t do it much better

On the other hand, attempts towards organizing the Web under large library-like classification schemes, at the measure of the Global Library itself, have emerged in the very first years of the Web. Yahoo! Directory, followed by its collaborative alternative Open Directory, the latter copied in hundreds of various avatars including Google Directory, have quickly met serious scalability issues, given the unprecedented size, variety and sheer instability of Web resources. Despite the bold tagline “Humans do it better” [7] the eventual failure of such attempts regarding their cost of maintenance compared to performance of automatic indexing has resulted in a general discredit of global Web classification tools and services, to the benefit of search engines algorithms. Nevertheless, a lot of vertical directories have kept flourishing, and are still alive.

A collateral damage of the Web directories paradigm failure could be to enforce in the average end user mind a vague notion that structured vocabularies are tools inherited from the dusty dark age of library classification, before Google was invented. On the other hand, the average end user is not aware of the differences between a thesaurus, a subject heading scheme, a classification, an index, a table of contents, a taxonomy, a directory … all avatars of vocabularies supposed to help her sorting out and finding things. For example the user of Wikipedia will be presented with a variety of “lists” and “categories” growing in an organic way, with no general plan, built on fuzzy guidelines by people not always well bred in the art of classification [8].

All in all, the average user experience regarding vocabularies on the Web is that they are as messy as the Web itself can be, with locally well ordered places in the middle of the general chaos. Instead of helping her sorting and finding stuff neatly and quickly, vocabularies on the Web will often simply add to the global knowledge soup [9], supporting at best some kind of lucky serendipity.

The raise and raise of *taxonomies*

Meanwhile in the closed worlds of enterprise information systems, the importance of structured and controlled vocabularies is more and more acknowledged and dealing with such vocabularies has become important enough to support the business model of a number of companies. In the enterprise realm, controlled vocabularies generally come under the generic name of “taxonomies”. In introduction to “The Accidental Taxonomist” [10] Heather Hedden writes:

> *For present-day information management, the term taxonomy is used both in the narrow sense, to mean a hierarchical classification or categorization system, and in the broad sense, in reference to any means of organizing concepts of knowledge. Some professionals do not even like to use the term, contending that it is too often ambiguous and frequently misused. Yet it has gained sufficient popularity, and a practical alternative term does not seem to exist.*

So, like it or not, your vocabularies, once published, and whatever their original structure and specific use are likely to be known in the wild as “taxonomies” [11].

Towards a semantic reconciliation

Opening the data silos, and the need for semantic interoperability

A new trend has been emerging in the recent years: the general opening of information systems. It has been widely acknowledged that duplicating and managing data in as many closed information silos and data bases had incredible costs, when a lot of data could be publicly available from servers “in the cloud” and consumed on demand. The distinction between the open Web and the enterprise information system becomes more and more fuzzy. SaaS (Software as a Service) has come of age, now followed by DaaS (Data as a Service).
Of course exchanging data between systems needs not only to share common data formats (such as XML), but also to agree on semantics of those data. *Semantic interoperability* has become a key issue, and for that matter it has become obvious that common reference vocabularies are needed, with formally declared semantics.

Semantic basics: dealing with ambiguity and co-reference

The Semantic Web stack is often presented as an artificial intelligence technology, the main point being the formal representation of things in languages based on description logics, such as OWL. Clearly this is the most advanced part of the stack, using complex constraints and rules to sort out inconsistencies and infer new knowledge. But before that, and certainly more important, Semantic Web languages and tools provide ways to deal with *ambiguity* and *co-reference*. To make it short, they allow to express explicitly if two things, bearing or not the same names, are the same or not, for example if two documents deal with the same subject, if they have the same author, if two resources are located in the same place etc.
In that respect, Semantic Web principles have always been closer to Library science than the so-called Web 1.0 infrastructure. Whereas Web 1.0 focuses on access to information resources through URL and non typed hypertext links, the Semantic Web deals first and over all with *identifying things resources are about*.

Identity of things, supporting disambiguation and co-reference, is ensured in the Semantic Web by universal identifiers (URIs) distinct from natural language terms used in library vocabularies. Nevertheless, the Semantic Web principles are closer to library science than to search engines logic. Aggregation of information in both universes is “concept-centric”, even if concepts are represented in the Semantic Web by URIs and formal description and in the library vocabularies by natural language terms. What was simply needed to have the two universes meet and work in synergy was a language to anchor library vocabularies to Web identifiers (URI).

SKOS: a vocabulary to rule them all?

The SKOS [12] language has been developed since 2005 through a fruitful dialogue between the libraries and Semantic Web communities, with the declared objective of allowing easy migration of legacy vocabularies A W3C recommendation since 2009, SKOS has immediately be used to support migration of legacy vocabularies, including RAMEAU[13], LCSH[14], AGROVOC[15], EUROVOC and many more. Most major vocabulary publishers are likely to follow those examples in the months and years to come.

SKOS has not been designed specifically for library vocabularies, but is intended to be usable for a large variety of Knowledge Organisation Systems, encompassing the "taxonomies" in the widest meaning presented above. Standard structured vocabularies such as thesauri can be translated to SKOS quite easily, but for some vocabularies with more complex features the expressivity of SKOS has to be extended.

Moreover, translating a vocabulary to SKOS format does not mean it is fine-tuned for efficient use in the Semantic Web. We will look in the following section at some practical issues, trying to identify what makes a vocabulary usable, and therefore how it can be improved for better efficiency.

Making library vocabularies ready for the Semantic Web

Integrating vocabularies has a clear added value for Semantic Web applications such as annotation, search and navigation tools, semantic extension of search etc. But it appears that some vocabularies are more efficient than others in this respect, those well suited for human usage not necessarily being the best fit for machines. Machines have not the subtle intelligence of human users, they need explicit definitions and disambiguation rules, the more so in particular for open environments calling for wide-scale interoperability. Machines have not far-fetched requirements. They only want semantic to be explicit, and they interpret the same syntax the same way, everywhere and always. A first and healthy side effect of porting a vocabulary to the Semantic Web will generally be to discover that it does not always mean what it's supposed to mean in the head of its authors and human users.

We now review a certain number of issues and lessons learned from the benchmark using RAMEAU, LCSH and SWD in the framework of TELplus. Highlighted are the issues clearly requiring evolution of vocabularies in order to make them actually and efficiently usable. They are presented in the form of good practice rules, of which some would apply to any data migration towards semantic applications, other are specific to library vocabularies.

1. Identify and describe non-ambiguous concepts

The Semantic Web basic assumptions regarding concepts, or for that matter absolutely any "thing" defined by a URIs are simple but strong.

1. The semantics (meaning of the URI) is context-independent. Wherever the URI is found, it means (denotes) the same thing. URIs are universal unique names.
2. The formal description of the concept, if possible obtained by dereferencing the URI through a Web protocol (most of the time, HTTP), conveys this semantics and makes the concept unambiguous.

In other words, the formal description of the concept should be explicit enough to make it clearly distinct from, or clearly identical to, any other concept defined by any other URI on the Web. This is quite a harsh requirement, but testing the published vocabulary in this regard is the first task of the potential semantic web user.

A policy of unique identifiers is needed, whereas general library vocabularies often rely on identification by names in context. Making distinct the concepts and the terms used to name them is critical, so the identifiers should be notations independent of labels. If the vocabulary management system uses internal unique keys for concepts, a good practice is to build URIs on those keys, such as for LCSH URIs presented below.

2. Explicit the semantics of your vocabulary syntax

When using vocabularies in library context, a lot of disambiguation can be performed by the human user. A term can be ambiguous in natural language, but no more so in a vocabulary context. “Table” is ambiguous in English (and French too) but it’s less so if it’s defined in the context of “Home furniture” or “Data base”. For Semantic Web applications the two concepts “Table (home furniture)” and “Table (data base)” should be identified as different concepts, by different URIs and linked in a formal way to their respective contexts. Those contexts have to be themselves formally identified and declared as either a broader or related concept, or a concept scheme (depending on the level of granularity). The simple indication of the context by a qualifier as in the thesaurus practice is not enough.

In a similar way to qualifiers, other specific syntactic structures hide implicit semantics which must be made explicit. The MARC format used by LCSH and RAMEAU contains syntactic constructions such as “Actors--Psychology” (http://id.loc.gov/authorities/sh85000748). Although the relation would seem obvious to human users, the formal description of this concept provides links to neither “Actors” (http://id.loc.gov/authorities/sh85000744) nor “Psychology” (http://id.loc.gov/authorities/sh85108459). For Semantic Web applications, this relation should be explicit to be usable.

A good side effect of working out such examples for the vocabulary manager is to force a reflection on the semantics underlying such a syntax, which she might have taken for granted for ages. For automatic migration to Semantic Web format, identical syntactic constructions should be interpreted and processed the same way. But it appears that finding out a general semantics for the above “--“ separator going beyond a general and vague “skos:related”, and valid throughout a huge vocabulary such as RAMEAU or LCSH, is a non-obvious task. Even if there are guidelines for vocabulary managers explaining in natural language what should be the correct use of each syntactic construct, generally the vocabulary management system has no way to enforce those guidelines, let alone assessing the consequences of incorrect use.

3. Check the actual transitivity of hierarchies

The hierarchical structure of vocabularies, likely to be translated into skos:broader or skos:narrower properties, will typically be used by semantic engines to perform extension or restriction of search to more generic or more specific concepts, in order to reduce noise or silence in results. Such extension is likely to follow the hierarchy in a transitive way. Even if there has been a lot of debate in SKOS on this issue, eventually closed by the decision that those properties are in general not transitive, many applications are likely to consider they are.
If one looks closely at large hierarchies such as LCSH or RAMEAU, it’s clear that semantic extension by transitivity is often valid “up to a point” or “locally”, so to speak. Since vocabulary managers have not necessarily a global vision of the vocabulary given its sheer size, some sort of “semantic drift” is likely to happen from top to bottom of the hierarchy, making semantic extension problematic. The following example taken from LCSH is not as extreme as it seems, we’ve found a lot of similar ones. In RAMEAU we even discovered some cycles spanning up to thirteen concepts, of course driving the semantic extension engines crazy.

Auxiliary sciences of history
.Civilization
..Learning and Scholarship
…Humanities
….Philosophy
…..Psychology
……Attention
…….Listening
……..Eavesdropping
……..Wiretapping

If one looks locally at this hierarchy, there is actually no flaw in any of the broader-narrower semantics. Each local relationship makes sense. But using it globally for semantics extension could lead to weird results.
There again, asking for explicit semantics will eventually lead the vocabulary manager to question the relevancy of such large hierarchies and the semantic drift they can entail, even inside the library system, not to speak about the open Web.

4. Keep it as small and simple as possible

The above examples tend to show well enough that large general vocabularies are not well adapted to semantic technologies, and should be better organized in vertical domains, where disambiguation is easier. An efficient vocabulary is either *shallow in meaning and wide in scope*, or *narrow in scope and deep in meaning*. Attempts to have both wide scope and deep meaning are leading to those huge vocabularies with hundreds of thousands of concepts, as difficult to maintain as to use. This is a main lesson learned the hard way, since those huge vocabularies are part of the legacy, and have been largely used for indexing. In order to leverage nevertheless the index legacy, mapping those large general vocabularies to smaller vertical ones, should be a priority task. Of course it will not be done in one day, but management tools built upon semantic web languages can help. We'll be back to this point in the last section.

5. Map to other vocabularies

Mapping to equivalent concepts in other vocabularies, singularly in multilingual environments, provides an important added value to the vocabulary. Even in monolingual environments, general vocabularies can be extended to more specialized ones, developed by independent parties. SKOS provides the expressivity for such mappings, and semantic search engines can leverage such mappings for semantic extension. In this respect, the work done in the MACS [16] alignment project is leading the way.
In particular, mapping general large vocabularies to smaller vocabularies as above defined will ease their use by navigation taxonomies providing simplified and customized views on library indexes and catalogues.

6. Concepts have a life cycle, but cool URIs don't change

URIs published for concepts in vocabularies should be robust over time, Semantic Web applications requiring that the URIs they use be as "cool" [17] as can be. Even in concepts in a vocabulary change, strict methods should be enforced to ensure the persistence of URIs and attached descriptions, even for obsolete concepts. Deprecation mechanisms can be put in place as well as redirection mechanisms using Semantic Web languages and protocols.

7. Expose and promote your Vocabulary as a Service

The Semantic Web is not only a pile of languages it's also an architecture of services. In order to be fully usable, the vocabulary should be available to applications in various forms, all supported by the Web architecture. Without entering the full details which are well described in excellent tutorials [18], let us sum up the main recommendations.

1. Each concept URI should support content negotiation, providing formal RDF description to machines and HTML for human consumption to regular browsers.
2. Make the vocabularies available in downloadable packages, one package by Concept Scheme.
3. Set up a SPARQL endpoint in order to allow users to extract parts of the vocabulary fitting specific requests.
4. Expose the content of vocabularies using for example the VoID [19] ontology.
5. Last but not least, promote your publication through the various forums of the Semantic Web community.

8. Use Semantic Web software for vocabulary management

Last but not least, development of software supporting the Semantic Web language brings about new tools to manage vocabularies, easing tasks such as consistency control, expressivity improvement using ontologies with more specific relationships than the simple hierarchical and associative relationships, handling of versioning, deprecation and evolution of concepts, mapping to other vocabularies, native import and export in RDF, built-in SPARQL endpoint.

In this respect, the European Community has set an example in adopting Semantic Web software stack to manage and publish EUROVOC and in the future many other vocabularies of the European Office of Publications. The solution has met high level management and workflow requirements for a multilingual vocabulary (over twenty languages), fine-grained management of terms, including synonyms, translations and acronyms, organization in micro-thesauri with different constraints, and of course publication in SKOS format, including versioning and time stamps.

Ongoing work and perspectives

We'll end up by a quick review of ongoing initiatives in the Semantic Web Library space. This is of course an open and non-exhaustive list, likely to grow in the next months and years.

Library of Congress Authorities and Vocabularies

The Library of Congress has led the path by starting publishing its vocabularies in formats conformant to Semantic Web standards and Linked Data best practices. The first set of vocabulary published in 2009 was LCSH, and others have been added in 2010. The automatic transformation of existing vocabularies in SKOS, as mentioned above, raises a certain number of issues, and this initiative has to be carefully assessed. But it is of course an incontrovertible reference, the id.loc.gov namespace is certainly here to stay, and any further

publication in the Library Semantic Web space should look at this space and connect to it as far as possible

The European Vocabularies

The preliminary studies made in the framework of TELplus have led to exchanges between the French National Library and its technical partners, in order to define the migration path of RAMEAU towards Semantic Web usability. We hope to see this work eventually extended to other partners in the Europeana and European Library framework. Multilingual and multicultural challenges are high, but it's a very promising space. We have mentioned above the initiatives of the European Office of Publications, and of course synergy and exchange of good practices between all those initiatives in Europe should eventually lead to the emergence of a rich multilingual, interconnected cloud of public vocabularies for access to Europe public assets of whichever nature: cultural, regulatory, economical.

W3C Library Linked Data Incubator Group

Started in May 2010, the W3C Library Linked Data Incubator Group[20] is chartered for one year to gather efforts in this domain, identify and promote best practices, and help the synergy between Library and Semantic Web lead both of them to their "full potential". Besides gathering main actors in the Librarian community, this group will also work in synergy with the ISO working group on the new Thesaurus standard (ISO 25964) and the Dublin Core Metadata Initiative.

Conclusion

We hope to have shown that now is the time for vocabulary managers to actively engage in a productive cooperation with the Semantic Web community, in order to see their precious legacy both improved and used in new ways. A framework for this task has been set up by the W3C, and quoting the charter of the Library Linked Data Incubator Group:

> A re-orientation in the library perspective on information interoperability is needed, building on existing Web architecture and standards, in order to bring this content to the Web. A lot of structured data is already available within library systems and could be released as Linked Data, using Semantic Web technologies. Cultural heritage institutions could be a major provider of authoritative datasets (persons, topics...) for the Linked Data Web.

The authoritative data sets are clearly here legacy vocabularies. The Semantic Web technology pile should allow librarians to manage them in a more efficient way, making them available through semantic service interfaces. In short, the expected evolution is to have vocabulary publishers open and extend their traditional usage to the Web at large, through a distributed and networked architecture of Vocabulary as a Service.

References and links

[1] OCLC WorldCat http://www.oclc.org/us/en/worldcat/default.htm
[2] W3C Semantic Web Activity http://www.w3.org/2001/sw/
[3] TELplus http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/
[4] Library of Congress Vocabularies and Authorities http://id.loc.gov/
[5] EUROVOC http://europa.eu/eurovoc/
[6] Search Engine History http://www.searchenginehistory.com/
[7] Humans Do It Better: Inside the Open Directory Project, Chris Sherman, ONLINE Mag, July 2000 http://www.onlinemag.net/ol2000/sherman7.html
[8] Wikipedia discussion about the "Uncategorized categories" list http://en.wikipedia.org/wiki/Wikipedia_talk:Special:UncategorizedCategories
[9] Representing Knowledge Soup in Language and Logic, John Sowa, 2002 http://www.jfsowa.com/talks/souprepr.htm
[10] The Accidental Taxonomist, by Heather Hedden, Information Today, May 2010, ISBN 978-1-57387-397-0 http://www.hedden-information.com/accidental-taxonomist.htm
[11] The Taxonomy Warehouse - http://www.taxonomywarehouse.com/
[12] Simple Knowledge Organization System http://www.w3.org/2004/02/skos/
[13] RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) http://rameau.bnf.fr/
[14] Library of Congress Subject Headings http://id.loc.gov/authorities#ConceptScheme
[15] AGROVOC Thesaurus http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[16] Integrating MACS initial data and new alignments into TEL framework http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/documents/TELplus_D3.4_04012010.pdf
[17] Cool URIs don't change, Tim Berners-Lee, 1998 http://www.w3.org/Provider/Style/URI
[18] How to Publish Linked Data on the Web, C.Bizer, R. Cyganiak, T. Heath. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
[19] Vocabulary of Interlinked Datasets http://semanticweb.org/wiki/VoiD
[20] Library Linked Data Incubator Group charter http://www.w3.org/2005/Incubator/lld/charter

About the author

Graduated from the Ecole Normale de l'Enseignement Technique (ENSET) in 1975, Bernard Vatant has taught mathematics in France's *Education Nationale* from 1975 to 1997. He further refocused his activity towards new technologies of knowledge representation, publication and sharing.

At the end of year 2000, he integrated the Mondeca team as a consultant in the domain of ontologies and knowledge representation languages. His experience in modelling and data migration has built up since that date, thanks to the diversity of Mondeca's customers and projects, of which scope encompass domains as various as scientific and medical terminologies, tourism and local government, or legal publication.

Bernard Vatant has been the representative for Mondeca in several working groups or standard bodies such as ISO (ISO 13250 Topic Maps standard, and current working groups about future ISO 25964 standard on Thesauri), or W3C (OWL, SKOS). He has chaired from 2001 to 2003 the OASIS Technical Committee on « Published Subjects ».

He's a known actor of the research on industrial use of Semantic Web technologies, and as such is a regular participant in Program Committees of dedicated conferences either in France, such as IC (Journées francophones d'Ingénierie des Connaissances) or abroad, such as LDOW (Linked Data on the Web), and has been a guest speaker to workshops organized by organisations such as INRIA in France or ISKO in UK.

His expertise is acknowledged in particular in the domain of modelling, migration and interoperability of legacy reference vocabularies, and in this respect has worked with institutions such as the Publications Office of the European Union (EUROVOC vocabulary), or the French national Library (BNF) in the framework of the TelPlus project (evolution and integration of RAMEAU vocabulary). He's since May 2010 an Invited Expert in the W3C Library Linked Data Incubator Group.