Qualitative and Quantitative Methods in Libraries (QQML) 4: 575-–583, 2015

Knowledge mapping and visualization as a common ground between librarianship and scholarly communication: Qualitative and quantitative methods for improving semantic categorization and retrieval

Liliana Gregori, Luca Losito, Paolo Sirito

Università Cattolica del Sacro Cuore, Milan Campus Library

Abstract: Current trends in the field of knowledge organization (KO) and a growing need for promoting scholarly communication are facing LIS professionals with new challenges as well as exciting opportunities.
Also academic libraries are urged to reshape their workflows and to focus on innovative technologies, in order to develop value-added services to its users: scholars, researchers and students.
According to such scenario, the Central Library of Università Cattolica del Sacro Cuore in Milan, launched in early 2012 a multi-year project, aimed to improve qualitative and quantitative aspects of semantic categorization and retrieval.
A ross domain ontology, based upon the *Protégé* framework, is currently being developed. When released, the system will provide patrons and staff with a visual interface, common to the three different tools currently used to manage bibliographic information:

- The library automation system (notably the cataloguing module)
- the Online Public Access Catalogue
- the institutional repository, named *PubliCatt*

The main expected benefits are:

- rationalizing and optimizing the cataloguing process
- facilitating and broadening access to all library resources
- making institutional research products more visible and citable

The paper will provide details about the project framework, as well as other notable deliverables, in particular:

- templates
- best practices
- lesson learned

Received: 21.10.2014 / Accepted: 1.3.2015 ISSN 2241-1925

Keywords: Scholarly communication; Ontologies**;** Knowledge organization**;** Information retrieval**;** Semantic categorization**;** Information visualization

1. Introduction

Since the publication of Wenger (1998), focused on aims and behaviors of communities of practice, the mission of academic libraries has dramatically changed: a new concept has been established in addressing the issue of relationship between library and faculty, by fostering a mutual exchange of contents, methodologies and resources.

The epistemic vision of communities of users and researchers, formulated by Haas (1992) and updated twenty years later by Cross (2012), has become the key issue for library services and academic planners.

Among the top trends for academic libraries listed by ACRL Research Planning and Review Committee (2012) we can find "*technology trends specific to libraries* (that) *include Web-scale discovery systems with enhancements such as discipline-scoped searching and customized widgets, community-source library management systems"*.

A survey to explore user experience activities at member libraries was conducted during the 2011 by the American Research Library (ARL): the methodology adopted and the collected results confirm the importance of holding an observatory related to usability issues and the challenge of representation and communication of academic discourse.

Information retrieval strategies are setting up a common ground between librarianship and scholarly communication, keeping the dialogue on these topics:

- how data are stored
- how information is organized
- how results can be represented

According to the aims of academic libraries worldwide (The Second Strategic Workshop on Information Retrieval in Lorne report (SWIRL), 15-17th Febr. 2012), a network of technologies must be developed in order to improve semantic categorization and to inform future practices in cataloging data, information searching habits, results retrieval.

2. Scenario analysis

Our purpose is not to browse through the history of the use of semantic web technologies in digital libraries. Rather, we managed to round up some clues from a selected benchmark of case studies, which were helpful and suitable to our needs. The need to draw a concept map supporting the shift between natural language and controlled terms is mostly perceived in domains that use a high specialized terminology as well as technical language, e.g. Biomedical Sciences, Humanities and Law.

First, we looked at some universities that have put into practice methodologies of user experience (UX) and interaction analysis in order to establish an observatory on behavior of staff and institutional users (University campus of

Berkeley, OE project 2011-2012; University of Washington, Internet-Based User Experience Lab 2010-2011; Stuttgart Media University, User Experience Research Group 2010-2011; Library User Experience July 2011; case studies: University of California, Massachussets Institute of technology, North Carolina State University, University of Michigan, Duke University, University of Virginia, Georgia Tech, University of Notre Dame, Rice University, Vanderbuilt University, University of Chicago, University of Kansas, Northwestern University).

A growing number of academic libraries have already tested the usefulness of an ontology-based information retrieval system for their users (e.g., Stanford University, University of Malaysia, Florida State Universities, Universities of Computer studies, Yangon).

Our context-specific challenge was to build some sort of a "double fold" system, enabling:

- from an internal perspective, easy semantic indexing for the library personnel, even junior ones. At the very same time, we tried to foster cross-domain subject cataloguing, in order to improve the overall efficiency rate of the system;
- from an external perspective, comfortable and efficient retrieval of information stored on University OPAC and Institutional Repository PubliCatt.

3. Project background

According to such scenario, the Central Library of Università Cattolica del Sacro Cuore in Milan, launched in early 2012 a multi-year project, aimed to improve qualitative and quantitative aspects of semantic categorization and retrieval.

From the very beginning, it was chosen to give a stricty functional and customer-oriented perspective to the project. Due to this reason, it was included into a broader initiative, called "Permanent observatory of the quality of service of the Central Library of the Catholic University".

The main aim of the observatory – established in 2011 – is to collect raw as well as structured information coming from both patrons and library personnel, in order to better target available resources and foster innovation, by means of well-targeted projects.

As a starting point, it was noted that the usage of subjects in querying the OPAC was pretty low (less than 10% of total searches), according to year 2012 statistics.

There was no great surprise about such primary evidence. It is in fact common wisdom that pretty complex searches tend to be neglected by patrons, so the analysis of the above mentioned data could have easily lea to a progressive dismissal of semantic cataloguing itself, but it was decided to better understand the overall phenomenon.

Therefore, the observatory team collected the feedback coming from the library colleagues involved in one-to-one reference transactions and they found that

Semantic searching was one of the most popular topics. The very same result came from the report of the ask@librarian service, the local virtual reference desk (VRD) available to Milan Campus students since 2009.
Further data drilling revealed that users were usually interested in semantic searching, but they did not feel comfortable with subject searching as a first step in their overall information retrieval strategy. In other words, they usually started their own queries by using keywords and referred to subjects only in a second part of the process. Unfortunately, this search strategy was not always performing as expected, as in some cases the subjects of the selected record were very specific and the hypertext navigation from one record to another one, bearing the same subject, offered to patrons some sort of circular results.
Curiously enough, in the very same period, a quite similar feeling came from the subject librarians engaged in semantic cataloguing. The team is currently composed of 11 people, with medium to high seniority and a degree-level instruction, usually in the very same topic where they are performing cataloguing activities.
Generally speaking, the felt that the usual time-honored workflow was definitely effective, but not equally efficient, due to the difficulty of keeping an adequate standard of quality, under the growing pressure from patrons.
Such concerns were escalated to and taken in due account by the Library Director, who charged the observatory of the quality of service with this project, too. Not surprisingly, it was pretty clear from the beginning that two above depicted situations could easily considered as two different sides of the same coin and that a comprehensive initiative could / should be started as soon as possible, in order to deliver added value to both involved parts:

- library patrons, getting more standardized and usable subjects, therefore revamping semantic search;
- library personnel, making use of a new, more performing cataloguing process, as well as innovative tools.

4. Methodology and tools

The project just entered in its second year and it is expected to be completed in early 2014.
The very first activity was to select the best methodologies and tools. It was therefore performed a careful review of current literature (as briefly exposed in chapter 1) as well as scrutiny of best practices at local and international level.
It was clear from the beginning that the role of technology as an enabler for the envisioned changes was crucial.
Therefore, it was requested that the original library team was integrated with a new professional role, built upon two different but synergic skill sets:

- document and workflow management
- semantic technology and information visualization

Thanks to the joint commitment of the Library Director and of top management of the Milan Campus of the University, who understood the innovation potential of this specific project and the enduring competitive advantage of a full set of new technology-enabled services, the project team was fully operational starting April 2012.
From a functional perspective, it was decided to build the system in a highly graphical manner, very different from the usual index-based systems.
It is therefore no surprise that it was clear from almost the beginning that ontologies could be the weapon of choice for both categories of potential users:

- patrons: students, professors and researchers
- library personnel: subject librarians and reference specialists

As a positive side effect, it is also worth mentioning that such graphics-intensive approach could lead, in the near future, to other broader projects – namely Linked Data – where usability and system interoperability are to be considered a must.
From a technology perspectives, it was chosen to follow the academic mainstream trend in choosing open source solutions. After a careful selection of all main available products and resources, it was chosen to select *Protégé* developed by Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine (*http://protege.stanford.edu).*
A cross domain ontology, based upon the *Protégé* framework, is currently being developed. When released, the system will provide patrons and staff with a visual interface, common to the three different tools currently used to manage bibliographic information:

- The library automation system (notably the cataloguing module)
- the Online Public Access Catalogue
- the institutional repository, named *PubliCatt*

At present time, the first step of the project has been completed in due time (deadline: March 2013) and the subject librarians can rely of a set of domain ontologies, enabling them to find out and pick subjects from a preselected set of controlled terms. In the next chapter more operational details will be provided.
The other two steps are currently on-going according to schedule, and some pilots are already available to selected users.
In this phase, particular attention is being devoted to usability issues, which is obviously connected to designing user-friendly and intuitive web interfaces. The next strategic decision to be taken in the near future will therefore regard the knowledge mapping and visualization tools to be deployed to end users (obviously in read only mode), which could be different from the *Protégé* interface currently made available to "power users" as subject librarians and reference specialists.

5. Project benefits and preliminary results

Our working experience is currently related to the first step of the project only, but we are using our library colleagues also as beta testers for the other project components, by simulating external access by patrons to Catholic University OPAC and the Institutional Repository PubliCatt.

As already mentioned before, the main expected benefit from the library side was related to rationalization and optimization of the semantic part of the cataloguing process.

Collected data shows that such objectives has been fully reached: starting from October 2012, date of release of the first domain ontology – Law and related subjects – the average processing time of an academic monograph (chosen document type for internal benchmarking purposes) has fallen by more than 35%.

The very same development pattern has been recorded for all areas which has been progressively involved in the initiative, with benefit ranging from 25% to 45% , depending from the subject areas and the seniority of the involved personnel.

Some more details, in order to better understanding the dynamics underlying a complex and knowledge intensive activity like semantic cataloguing:

- Colleagues involved in Humanities got the best benefits from the streamlining of the subjects, which came as a welcome side effect of ontology building
- Junior colleagues got a significant improvement of their learning curve (approximately an half of the usual training time)

According to such figures and taking into due account the evidence collected from the field, we can argue that the subject areas like Humanities where semantic complexity is very high and subject picking mostly discretional (as not directly based upon keywords) get a clear benefit from such approach, provided that the whole team have been committed to the project from the very beginning, that is from the ontology building phase.

Interesting enough, one of the reason of the success project also from the senior cataloguers perspective (potentially more traditionally-oriented and resilient to change) was the idea of leaving some sort of heritage, in the form of a domain ontology, to younger generations of library professionals.

Such approach has been promoted by the Library Director, as a tangible sign of respect for all library staff, whose commitment in the project have been essential in the first phase and whose deliverable (see next and final paragraph) are currently used for the other two project phases.

From the users' perspective, we expected that – after one full year of subject loading (that is: October 2013) the number of average records retrieved from the OPAC for a single semantic string should range from 5 up to 20 items, avoiding therefore too generic or too specific terms.

The perceived value for the users will rise proportionally with the number of records produced with the new methodology and we are thinking about the

possibility to make some sort of retrospective updating of the subjects, in order to make them fully searchable with the new system.
Finally, it is worth mentioning that the reference specialists are already testing the acceptance of the new methodology through the regular sessions they hold with patrons (one to one interviews as well as small groups seminars) and collecting a structured feedback. Obviously, the new graphical interface for the OPAC (expected early 2014) will shift the overall user experience to a definitely higher level and unleash the full potential of the ontology driven approach to semantic cataloguing.

6. Best practices and lesson learned

Although it is difficult to give full details, due to space constraints, about specific deliverables, like templates, which require a different degree of analysis, we can anyway focus on other interesting topics:

- best practices
- lesson learned

About best practices, from the technology side of the project, we obviously acknowledge full credits to the team who developed and to the whole community who is supporting *Protégé.* As mentioned before, this work was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.
We have been fully satisfied by the functionalities provided by *Protégé* in the ontology building phase, which involved only the project team (three people) and the subject librarians, but we have not decided yet which kind of tools we will be using, in order to implement the knowledge visualization interfaces which will be released – in read-only mode – to the users of both OPAC and PubliCatt.
A software selection is currently on-going, taken into account a full range of solutions, including commercial ones, as robustness and scalability will be two key issues, while dealing with several concurrent users on multiple platforms, including mobile ones.
A special attention has been given to the user experience issues, which are the distinctive mark of the Observatory on the Quality of Service of the Catholic University. This led to a quite complex and time consuming fine crafting of the user interface finally released to subject librarians. The results were anyway fully worth the effort.
Regarding lesson learned, they can be synthetized in three points:

- Do not assume that subject librarians have necessarily a strong background in information architecture. The project team devoted a significant deal of time in training library colleagues (although experienced and proficient in their own subject domains) in ontology building.

- Quantitative aspects of knowledge mapping may be tricky. Even with this careful preliminary training, the drafts results of ontology crafting were very heterogeneous and sometimes unsatisfactory, so to need a second round of training. The main issue was finding out a common and reasonable number of nodes (that is: controlled terms) constituting each domain ontology. Finally, it was agreed that such number should range between 500 for simpler ontologies (such as Sociology) and 2.000 for more complex ones (like Law and Economics)
- Cross domain issues have to be managed at an higher level, but are not critical by themselves. Obviously we had to take into account that a certain number of concept (like federalism) are to be included into different domain ontologies (Law, Economics, Political Science), but we realized that subject librarians were not mostly interested in such cross domain issues. Therefore, the project team drafted an high level ontology which was submitted for validation to the Library Director and then implemented through a series of cross references in *Protégé.*

References

Cross, M. K. D., (2013). Rethinking epistemic communities twenty years later. *Review of international studies*, 39, 137 – 160.

Simons, N. and Richardson, J., (2012). New Roles, New Responsibilities: Examining Training Needs of Repository Staff. *Journal of Librarianship and Scholarly Communication*, 1(2):eP1051.

Aung, N. N. and Naing, T. T., (2011). Sports Information Retrieval with Semantic Relationships of Ontology, *3rd International Conference on Information and Financial Engineering IPEDR*, Vol.2. IACSIT Press, Singapore.

Freire, N., Borbinha, J. and Calado, P., (2011). A Language Independent Approach for Aligning Subject Heading Systems with Geographic Ontologies. Proceedings of the *International Conference on Dublin Core and Metadata Applications*, North America, 0, sep. 2011.

Grimm, S. and Wissmann, J., (2011). Elimination of redundancy in Ontologies. *The Semantic Web: Research and Applications*. Proceedings of the *8th Extended Semantic Web Conference, Heraklion, Crete, Greece, May 29-June 2, 2011*. Springer, Berlin, 260 – 274.

Tsakonas, G. and Papatheodorou C., (2011). An Ontological Representation of the Digital Library Evaluation Domain. *Journal of the American Society for Information Science and Technology*. Vol. 62, No. 8, 1577 – 1593.

Zaid, N. M. and Lau, S. K., (2011). Development of Ontology Information Retrieval System for Novice Researchers in Malaysia. *Journal of Software and Systems Development*, 2011, 1 – 11.

Fernandez-Lopez, M. and Corcho, O., (2010). *Ontological Engineering*, Springer, Berlin.

Mason, D., (2010). Progressive concepts for semantic web evolution: applications and developments. *Online Information Review*, Vol.35, No. 1, 166 – 167.

Papadakis, I., Kyprianos, K., Mavropodi, R. and Stefanidakis, M., (2009). Subject-based

Information Retrieval within Digital Libraries Employing LCSHs, *D-Lib Magazine*, Vol.15, No. 9/10.

Papadakis, I., Stefanidakis, M. and Tzali, A. (2008). Visualizing OPAC subject headings, *Library High Tech*, Vol.26, No. 1, 19 – 23.

Park, J.-H. (2008). The Relationship between Scholarly Communication and Science and Technology Studies (STS). *Journal of Scholarly Publishing* , Vol.39, No. 3, 257 – 273.

De Keyser, P., (2007). *Indexing: from thesauri to the Semantic Web*, Chandos, Oxford.

Taniar, D. and Rahayu, J.W., (2006). Web Semantics and Ontology, Idea Group, Hershey-London.

Gnoli, C., Marino, V. and Rosati, L., (2006). *Organizzare la conoscenza. Dalle biblioteche all'architettura dell'informazione per il web*, HOPS Tecniche nuove, Milano.

Yi, M., Burnett, K. and Burnett, G., (2006). *User Performance Using an Ontology-Driven, Information Retrieval (ONTOIR) System*, Florida State University Tallahassee FL.

Taylor, A.G., (2004). *The organization of information*, Libraries Unlimited, Westport CT.

Gajo, M.G., Mariani, A. and Villani, M.G., (2003). *Semantic Web and libraries : 26th Library Systems Seminar, Rome, 17-19 April 2002*. Biblioteca nazionale centrale di Roma, Roma.

Berners-Lee, T., Hendler, J. and Lassila, O., (2001). The semantic web. *Scientific American*, Vol.5, 34 – 43.

Wenger, E., (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, Cambridge.

Haas, P.M., (1992). Introduction: epistemic communities and international policy coordination. *International Organization*, Vol.46, No. 1, 1 – 35