Organizational Challenges of the Semantic Web in Digital Libraries

Bendik Bygstad[*], Gheorghita Ghinea[*+] and Geir-Tore Klaebo[+]

[*]*Norwegian School of Information Technology, Oslo, Norway*
[+]*School of Information System, Computing and Mathematics, Brunel University, U.K*
e-mail: bygben@nith.no; george.ghinea@brunel.ac.uk; geir.tore@minar.no

Abstract

The Semantic Web initiative holds large promises for the future. There is, however, a considerable gap in Semantic Web research between the contributions in the technological field and the research done in the organizational field. This paper examines, from a socio-technical point of view the impact of Semantic Web technology on the strategic, organizational and technological levels.

Building on a comprehensive case study at the National Library in Norway our findings indicate that the highest impact will be at the organizational level. The reason is mainly because inter-organizational and cross-organizational structures have to be established to address the problems of ontology engineering, and a development framework for ontology engineering in digital libraries must be examined.

1. Introduction

For organizations with large amounts of non-numerical data the Semantic Web initiative holds promises for the future [2]. The past decade gave us not only the Semantic Web and eXtensible Markup Language (XML), but also such tools as the Resource Description Framework (RDF) and ontology languages [12]. All together, these concepts and tools provide us with a powerful environment to escape the lexical doldrums of the World Wide Web.

Organizations, however, are socio-technical systems; to work they need both technology and people. The progress of technologies has not been matched by an equal understanding of the organizational issues and challenges associated to the Semantic Web. This applies in particular to one of the largest potential users of this framework, namely large libraries. With the notable exception of a contribution from Kim and Biehl [7], little research has been done in the area of organizational implications of implementing Semantic Web in digital libraries. Some of the research done by Klischewski [8] does have some relevance to digital libraries and his agenda for further research suggests focusing on cross-organizational adoption of the Semantic Web and the ability to set up and manage socio-technical infrastructures. This agenda may also be appropriate to research in digital libraries; however his focus on e-Government as a basis for his research represents a limitation because of the specific definition and use of ontologies in e-Government, and the structure of the organization which is not necessarily transferable to digital libraries. This suggests that an investigation of the strategic and organizational implications implementing Semantic Web in Digital libraries would be appropriate, and represents the focus of the current paper.

Our research question is*: "what are the strategic, organizational and technological impacts of the Semantic Web on large digital libraries?"* The rest of the paper is structured as follows. In section 2 we discuss the opportunities for large libraries presented by the Semantic Web. In section 3 we describe our research approach and present our case, the National Library of Norway. We discuss our findings in section 4, and conclude briefly in section 5

2. The Semantic Web and Digital Libraries

One of the main challenges in information retrieval from the World Wide Web is how to determine what is most relevant for a user's request. The basic elements in the most common and widespread search engines are purely statistical calculations and referential frequency from other sites. The need for a more semantic oriented information representation on the web has been acknowledged for some years, and considerable research activities have been performed in the field. The benefit of the semantic approach based on ontologies is that it gives us a framework for searching and browsing information objects on the web and gives more relevance and accuracy to search processes. It will also enable us to put machine-

978-1-4244-3397-1/08/$25.00 ©2008 IEEE

readable data on the Web, data which can be processed by automated tools as well as people [2].

The Semantic Web and semantic technologies are focused around knowledge representation through ontologies. An ontology is defined as a formal representation of the knowledge in a specific domain i.e. what exists can be formally represented by ontologies [3]. In the context of the Semantic Web, ontologies can be queried and updated, both by computer and human users, to explicitly represent objects and concepts that exists in some context, together with the relationship that holds among them. Adams related ontologies to the librarian taxonomy term, calling both hierarchies of structured vocabularies [1]. However, ontologies also include a set of semantic rules which is used to infer knowledge from a structured hierarchy of information, thus giving the complete structure a semantic meaning [5].

Digital libraries

A digital library is a "particular kind of information system and consists of a set of components, typically a collection (or collections), a computer system offering diverse services on the collection (a technical infrastructure), people, and the environment (or usage), for which the system is built" [4]. Digital libraries today are mainly repositories of digitized documents and if they are to become repositories of knowledge, and thereby represent a strategic asset for the organization, semantic annotation has to be connected to the digital content and semantic meaning has to be drawn from the documents. However many digital libraries hold not only digitized documents but also other several different media types such as digitized sound, film and pictures. The key challenges for digital libraries were already in the mid nineties identified as [9]:

Interoperability: The ability of digital libraries to share and relate information between different types of digital content across heterogeneous platforms.

Description of objects and repositories: The need for a commonly accepted naming convention in the description of objects and repositories to facilitate search and information retrieval from different distributed resources.

The management of storage and collection of information: The ability to store, index and retrieve non-textual and multimedia content.

User interface and human-computer interaction. How information is visualized and presented to the user, and how a user is to navigate in large information repositories.

The Semantic Web contributes to a solution to these challenges [12]. However, the challenges are not limited to the information structuring, but relate also to strategy and organization. For instance, in the context of digital libraries, search capabilities are often limited by the general search technology used by traditional search engines, and are often off-the-shelf software from commercial search companies like Google and Fast Search and Transfer. Digital libraries, however, are usually more complex, with multiple media-types involved and meta-data stored in several databases, and the interoperability challenge is very much at the forefront in these libraries. Nevertheless, the search capabilities in these digital libraries are limited in much the same way as for more rudimentary digital libraries because of the lack of both semantic annotation and ontologies. Indeed, one of the biggest challenges for organizations posed by the Semantic Web is no doubt the building and maintenance of ontologies. This challenge is not only technical, as our case study demonstrates.

3. Case Study

For the investigation of the research question we chose a qualitative and interpretive approach (Walsham, 1993), conducting a case study. The selected case was a large scale digitizing project in the Norwegian National Library (NL). The NL has decided to digitize all its material covering 15 media types, and make this material available for users on the WWW. It has carried out a pilot project during the last two years to gain experience and is commencing a second phase of the project where new material types are to be digitized and made available.

The case study builds on two major sources of information: a) the documentation of the Digitizing project in NL and b) interviews with nine different stakeholders at three levels of NL's organization during the summer of 2007. The informants were selected as follows: At each level three informants were interviewed. Top managers were interviewed on strategy, middle managers and librarians were interviewed regarding organizational issues and ICT-professionals were interviewed on technology issues. The interviews were conducted in a semi-structured manner, and were taped for analysis of data.

The data were analyzed at the three organizational levels, drawing on the nine interviews and on the substantial amount of project documentation. Findings from the interviews were structured according to these topics, and analyzed according to the relative impact each area has on the three levels of the research; strategy, organization and technology,

The Digitizing Project

The Digitizing Project was initially born as a result of a strategic process in NL focusing on The Digital National Library [11]. NL had been digitizing different media types for more than a decade before the advent of the digitizing project. In 2005 NL's digital collection consisted of more than 50.000 hours of radio programs, 200.000 photos and more than 200.000 newspapers issues. The NL had already established a digitizing production line for these material types, and has gained considerable experience in the field. The new elements in the Digitizing Project was an additional 11 material types to be added to the production line, and all material types were to be made available through a generic user interface using traditional search engine technology. A pilot project was initialized during the spring of 2006 adding books and magazines to the digitizing production line, and a user interface was designed to make the digitized material available for users. NLs digital library was officially opened in April 2007. See http://www.nb.no.

Two areas were criticized; a) the lack of semantic context in the digitized material and b) the lack of tools and instruments for users to interact and to add domain knowledge to the information objects. In its defence, NL stated that the released product was a Beta-product, and that the purpose of the pilot project was to gain experience in the field.

When the second pilot project is completed the ambition is to commit the base organization to take over the responsibility for the process because the time span of the total project is ten years, and it is not suitable to organize this process as a single project. The process will therefore be organized as a program, and several smaller projects will be initiated during the next ten years, hopefully resulting in a total digitized collection at NL in 2018.

Analysis and Discussion

One observation made during the interviews was that the participants had different perceptions of the concept of the Semantic Web. However, the need for more semantics in information retrieval was expressed by most of the participants, and the knowledge of topic-catalogues/topic-maps helped in the understanding of the need for ontologies in a semantic paradigm. The findings are summed up in Table 1.

Level	Main challenges
Strategic	Top-level support and funding depends on a broad understanding of the key issues Strategic semantic meta-data issues involving the organization have to be decided
Organi-zational	Ontology production and maintenance needs to be organized and managed Semantic annotation of information objects is an entirely new paradigm, in need of new competence and management. A new semantic based meta-data strategy is needed Cross-organizational coordination is needed
Technical	Tools used today are inadequate in a Semantic Web paradigm Open standards are mandatory

Table 1: Semantic Web impact

Strategic level

Arguably, libraries have made their first critical strategic decision regarding the Semantic Web when they decide to start the process of digitizing their information objects. The second strategic decision is how and what kind of meta-data is to be produced and how this is going to be organized [10]. The first decision is technology focused, while the second is socio-technical and cross-organizationally focused.

The findings at the strategic level indicate that the introduction of Semantic Web technologies will have high impact on top management involvement and resource availability. This is mainly based on the interviews and the experience from the digitizing pilot project: if Semantic Web technology is to be implemented the entire organization must support the initiative. On resource availability the indicator is most likely affected by the resource availability on the current digitizing pilot project, and the awareness that even semi-automated semantic annotation will be highly resource consuming.

The interviews indicated that the current organizational structure is built to support meta-data production, and therefore there is no need for re-organizing because of the introduction of the Semantic Web. However the socio-technical infrastructure needed to build and maintain ontologies might demand some changes in the organizational structure. Ontology building will be an inter- and cross-organizational effort, especially when there is different material types involved which are handled by different parts of the organization. This is, however, mainly an organizational challenge.

Organizational level

At the organizational level the impact is perceived as generally high. Some informants describe the magnitude of the organizational challenges facing NL of the Semantic Web as overwhelming. The basis for this is the issues of ontology production and maintenance, and the strong focus on semantic meta-data production, all of which have to be established. For the digitizing pilot project this would probably involve a total re-thinking of the role of semantic meta-data and how meta-data in general is produced, and the introduction of semi-automated Knowledge Discovery (KD) systems is likely to be the consequence. One middle manager said:

> *"The meta-data production in the digitizing pilot project is primarily a registration of meta-data for preservation purposes. Semantic annotation of the information objects would be an entirely new paradigm for the organization."*

Regarding organizational involvement the interviews indicate that the Semantic Web will have a high impact on this area, as illustrated in the above quotation from another middle manager:

> "*The organizaton have to be more involved in the Digitizing pilot project both in the planning phase and the operational phase of every new type of information objects set in to production."*

A high degree of organizational change for digital libraries transforming to Semantic Web technology this involvement is seen as crucial, because the knowledge in ontology engineering and maintenance and in the semantic annotation of documents will be an inter- and cross-organizational knowledge. This is in line with Klischewski's [8] call for cross-organizational infrastructures. The interviews also indicate concerns related to the present meta-data strategy, and the perception that a new semantic based meta-data strategy is needed in a Semantic Web paradigm.

On organizational consequences the interviews indicates the need for cross-organizational structures. On the other hand some informants express a certain fatigue in the organization because of the re-organizations process the organization recently have been through, wanting to solve the matter within the existing organizational structures. The interviews also indicate that the resource availability is a high-score area. This was a hot topic in the on-going Digitizing pilot project, and there is an expressed concern for the general resource availability in the organization.

Any new information system will have some impact on the organizational level when introduced and implemented. This is also the case for the Semantic Web but informants highlight that when this technology is implemented in a digital library the organizational impact will be universal because the main organizational activities in this organization is the digitizing of information objects and meta-data production [6]. In a Semantic Web paradigm ontology engineering and maintenance will be an additional task which might possibly force the organization into a more inter-organizational and cross-organizational structure. In this situation the question about organizational readiness and structural changes to meet the challenges from the Semantic Web is highly adequate.

Technological level

At the technology level the interviews indicate that the overall impact from Semantic Web technology is considered to be low or medium. NL has been digitizing content for over a decade, and in such areas as the digitizing of the non-textual material types, they have competent and experienced personnel. The impact Semantic Web technology will have on the Digitizing project is therefore regarded by the informants as low, probably because of the high confidence in their own technological capabilities. However, the organization lacks first-hand experience in ontology engineering, and the software tools available for this purpose are not known in the organization. Although there are discussions about semantic structure, RDF is not mentioned as an applicable standard for creating these structures. However the XML experience and competence in the organization is high and this would probably reduce the impact at this level.

Moreover, the Impact from the Semantic Web on technology choices is perceived as low, because the technology platforms in use at NL today are mainly based on open standards and will support the implementation of Semantic Web technology. There is, however, a perception that the tools used today are inadequate in a Semantic Web paradigm. This may be somewhat surprising (given the rich functionality of the current technical environment), but it may be explained by the lack of experience in ontology engineering.

5. Conclusion

There is a considerable gap in the Semantic Web research field between the research done in the technological field and the research done in socio-technical field. This paper is a contribution to understand and to explain how and digital libraries are affected when introducing Semantic Web technology. The research documented in this paper has investigated the different strategic, organizational and technological impacts of the Semantic Web on a large national digital library. We offer three conclusions from this research:

On the strategic level the impact from Semantic Web Technology in digital libraries is moderate. This conclusion is based on the fact that the strategic technology choices have already been made when the library have decided to move from a traditional library towards a digital library. However, strategic semantic meta-data issues involving the organization have usually not been decided because of the lack of an Ontology engineering development framework and a semantic meta-data annotation tool which are able to reduce the cost of the meta-data annotation. This represents an unknown risk factor at the strategic level. It also illustrated that top-level support and funding depends on a broad understanding of the key issues.

On the organizational level the impact from Semantic Web Technology will be high because both the ontology engineering process and the semantic meta-data production will affect the entire organization both on the inter-organizational and the cross-organizational level. For public digital libraries this probably will call for a more open policy towards user groups to manage the process of ontology engineering in a proper manner.

On the technology level the impact will be relatively low because the technology needed to support the traditional digital library already is in place and the leap from this technology to Semantic Web technology is not a giant one.

We acknowledge that these conclusions may be constrained by the fact that a single case approach is chosen, and the fact that the knowledge of Semantic Web technology among the interview subjects was limited. The overall conclusion, however, that the largest impact of the Semantic Web is at the organizational level, is supported by other research. This conclusion might be assumed to be universal to the digital library community because they are all facing the same challenges regarding semantic meta-data production and ontology engineering and maintenance, regardless of organizational connection (public/enterprise) and regardless of language and geographic location.

References

[1] Adams, K. The Semantic Web: Differentiating between taxonomies and Ontologies, 2002, online, 26, 4, pp 20-23.

[2] Berners-Lee, T., Hendler, J. and Lassila, O.The Semantic Web: a new form of web content that is meaningful to computers will unleash a new revolution of possibilities, Scientific American, 2001,5, 1.

[3] Ferran, N., Mor, E. and Minguillón, J. Towards personalization in digital libraries through Ontologies, Library Management, 2003, 26,4/5, pp 206 – 217

[4] Fuhr, N., G.Tsakonas, T.Aalberg,M.Agosti, P.Hansen, S. Kapidakis, C. Klas, L. Kovács, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters and I. Sølvberg. Evaluation of digital libraries, International Journal of Digital Libraries, 2007, 8:21–38

[5] Gruber, T. A translation approach to portable ontology specifications. International Journal of Knowledge Acquisition and Knowledge-Based system, 1995, 5, 2, pp 199-220

[6] Hauknes, R. Project report from the DigitALT project, National library of Norway internal document, 2006, NB.

[7] Kim, H. M.and Biehl, M. Exploiting the Small-Worlds of the Semantic Web to Connect Heterogeneous, Local Ontologies, Information Technology and Management, 2005, 6, pp 89 – 96

[8] Klischewski, R. Ontologies for e-document management in public administration, Business Process Management Journal, 2006, 12, 1, pp 34 - 47

[9] Lynch, C. and Garcia-Molina, H. ITTA Digital Libraries Workshop, 1995, available at www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html

[10] Lytras, M., Sicilia, M-A., Davies, J. and Kashyap, V. Digital Libraries in the knowledge era: Knowledge management and Semantic Web technologies. Library Mangement, 2005, 26, 4/5, pp 170-175.

[11] NB. Digitizing of the National Library's collection, Project overview, National library of Norway internal document, 2006.

[12] Warren, P. and Alsmeyer, D. Applying semantic technology to a digital library: a case study, Library Management, 2005, 26,4/5, pp 196 – 205.