Campbell, D. G., Brudin, M., MacLean, G., and Baird, C. (2007). Everything old is new again: Finding a place for knowledge structures in a satisficing world. *Proceedings of the North American Society for Knowledge Organization.* Vol. 1. Available: http://dlist.sir.arizona.edu/1896/

D. Grant Campbell, Michael Brundin, Graham MacLean and Catherine Baird
University of Western Ontario, London, Ontario, Canada

Everything old is new again: Finding a place for knowledge structures in a satisficing world

Abstract: The authors use an exploratory project involving Web resources related to Alzheimer's Disease to explore ways in RDF metadata can more effectively translate the virtues of the traditional vertical file to a Web environment form using Semantic Web descriptive standards. In so doing, they argue against the separation of "bibliographic control" from the socially-embedded institutional practices of reference work, collection development, and the management of information ephemera. Libraries of the future will use specific Web technologies that lend themselves to sophisticated and rigorous knowledge structures, and link them with librarians' skills in information harvesting and evaluation.

1. Introduction

This paper discusses the relationship between traditional libraries and the World Wide Web Consortium's standards for a Semantic Web. Both of these entities have acquired a reputation for being out of step and out of date in a Web 2.0 world which is hostile to large, intricate, rigorously-defined standards of information description and classification. We will use an ongoing exploratory project involving RDF metadata to describe a potential alliance of these unfashionable entities, in which traditional information institutions such as libraries could adopt Semantic Web technologies to redefine their services.

2. Libraries and the Semantic Web

Traditional libraries frequently suffer from a disjunction between an ambition to provide a useful and vibrant public space and an adherence to information systems that are incomprehensible to modern users. While public and academic libraries retain some cachet as valuable repositories and providers of important documents and programs, the library catalogue in most of its OPAC manifestations appears more and more of an outlier to students and citizens accustomed to Google and other Web-based services (Fast & Campbell, 2005). While the ubiquity of the MARC standard, together with the presence of enormous bibliographic utilities such as OCLC provide a powerful incentive for libraries to remain linked to their old standards such as *AACR2R, LCC, DDC, LCSH* and *UDC*, the complexity of these tools, together with their painfully slow rate of change, cause frustration in the library community, and bewilderment outside that community (Coyle & Hillman, 2007).

To call the Semantic Web "old-fashioned" might raise more eyebrows. Although the Semantic Web has been under development since the mid-1990s, the possibilities that it strives for continue to intrigue and excite information professionals and academics who hope that it will

one day reach critical mass. The prospect of resource descriptions encoded in the Resource Description Framework (RDF), linked together with ontologies encoded in languages such as the Working Language for Ontologies (OWL) continue to raise glamorous possibilities:

- consistent access to database information in the Deep Web by using XML and RDF as a lingua franca;
- machine-understandable data that can enable intelligent agents to make rudimentary inferences;
- rich and effective cross-domain resource discovery through common metadata and metadata encoding standards (Passin, 2004, 5).

Unfortunately, critical mass has not yet been reached, and some have grown tired of waiting. Clay Shirky argues that the Semantic Web is good for syllogisms and not much else (2003), and comes down in favour of looser, less-structured systems of organization like user-centered tagging services, on the grounds that "they don't recreate the structured, hierarchical categorization so often forced onto us by our physical systems" (2005). And Tim Bray offers a discouraging indictment of the Semantic Web's success rate:

> RDF is well into middle age as standards go, and it hasn't exactly turned the world inside out. This despite fierce backing from Tim Berners-Lee, who sees RDF as a key foundation component for the Semantic Web. In fact, as far as I know I've never used an RDF application, nor do I know of any that make me want to use them. (2003)

Proponents of both libraries and the Semantic Web, therefore, hear the same advice fired at them time and again: relax the insistence on detailed, rigorous and painstaking descriptions, and adopt a satisficing approach. Libraries should stop trying to organize everything according to their persnickety standards, and instead learn to adapt to a world in which information is largely ephemeral and task-related, with a very short shelf life. Similarly, the Semantic Web is considered far too complicated, given how easy it is to develop tagging systems that, while far from perfect, are often perceived as good enough.

This is depressing for both communities to hear. Libraries have always had a love-hate relationship with ephemera: those endless piles of pamphlets, course calendars, theater brochures, clippings of local interest and broadsides that are often useful in answering reference questions, but are difficult to find because their temporary usefulness and small size makes them prohibitively expensive to catalogue. To work in an environment consisting of nothing *but* ephemera would be daunting indeed. Similarly, the Semantic Web standards are developed by those who stubbornly continue to believe that there is a place for highly-structured information on the Web, and are unwilling to believe that such environments are impossible or impractical. Libraries and the Semantic Web, then, have a common, if unglamorous cause: to insist on the importance of careful descriptions of trusted and reliable information in a world which seems hostile to such care. Is there any way they can link together?

3. The matching process

To answer this question, we went back to F.W. Lancaster's famous paradigm of information retrieval as a matching process between two balanced processes of analysis and translation (1986). In knowledge organization, we are familiar with the process whereby a newly-acquired document is analyzed to extract a statement of its content, which is then translated into the language of the system, including a classification system and a controlled vocabulary. We sometimes forget that the success of this analysis-translation process, in Lancaster's view, depends on an equally important process, whereby the user articulates an information need which is translated into the system language through the mediation of either a human intermediary, in the form of a reference librarian, or the information system itself in the form of syndetic references and orderly displays of authors, works and editions. Only when these dual processes match, Lancaster claims, can the information retrieval process be successful. If that is the case, the experience of collection and reference librarians in assessing resources and answering information queries could provide insights into the problem of adapting current library systems. They could also provide insights into the Semantic Web, since the services they offer—acquiring reliable information sources and mediating different articulations of an information need—are very close to the value-added features of a Semantic Web.

4. Early directories

In the early days of the World Wide Web, libraries and librarians participated enthusiastically in the creation of Web resources. Large clearinghouses, such as the Internet Public Library (http://www.ipl.org/), organized links to vast numbers of resources according to spatial or intellectual analogues to physical libraries. Some, like the BUBL service (http://bubl.ac.uk/), used Dewey numbers as a means of classifying the resources (see Figure 1).

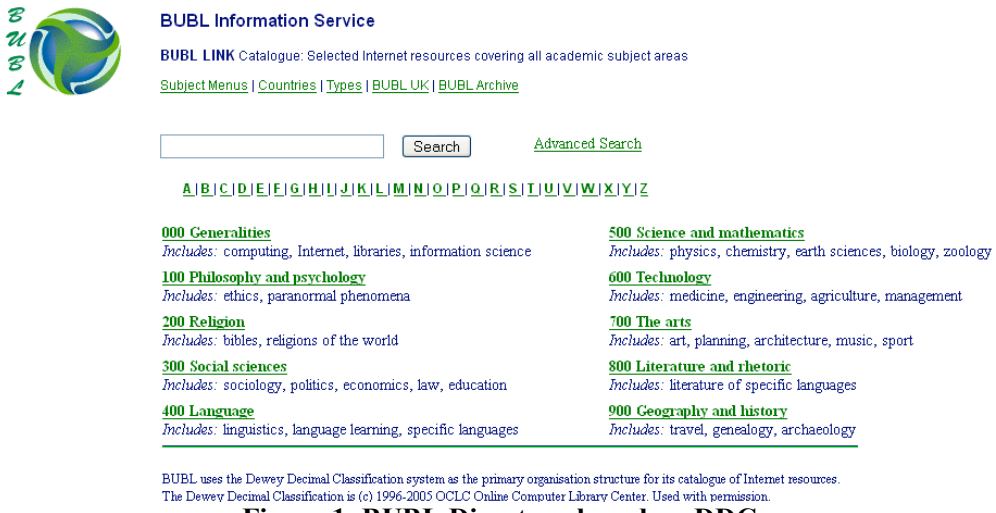

Figure 1. BUBL Directory based on DDC

As time went on, such directories gave way to those which echoed the structure and appearance of Web directories such as the Yahoo! and Open Project directories. The most successful of these, perhaps, was the Librarians' Internet Index (http://www.lii.org/), designed to imitate the Yahoo practice of showing categories with subheadings beneath each. Such directories frequently follow a discouraging track: they start small, then grow, then expire due to staff exhaustion, funding cuts or a change of ownership. While a few relics of that era remain in the form of the Open Project Directory and BUBL, most have collapsed under the weight of work and money required to maintain them.

Libraries have since filled the gap with their own resource lists. And some in the profession have made the connection between their Web resource lists and vertical files: file systems maintained for library patrons or staff for collecting ephemeral print resources relevant to topics of current or local interest (Colburn, 1997; Neuhaus, 1997; Thomsen, 1997).

While vertical files have disappeared from many libraries since the advent of the Web, in their heyday they possessed certain advantages that warrant consideration. First, since they could be used to collect ephemeral items such as pamphlets and magazine articles, they were useful for collecting information on topics at a level that frequently by-passed conventional publishing process. Early consumer health information services, for instance, often made use of vertical files for items that provided health information in language for non-professionals. Second, vertical file systems were typically created and maintained by reference staff in response to patterns noticed in patron inquiries, thereby rooting them firmly in the analysis and translation of reference interviews. And finally, vertical file systems were typically local in focus and modest in scope, with almost no likelihood of expanding beyond a staff's budget or resources.

5. The Resource Description Framework

The Resource Description Framework, as originally conceived by Tim Berners-Lee and the World Wide Web Consortium, is a method for making statements about resources in a structure that distantly resembles the syntax of a grammatical sentence, containing a subject (the resource being described, typically expressed as a URI), a predicate (which names a particular attribute of the resource), and an object (the value of the attribute) (see Figure 2). This is very similar to the cataloguing process, which involves making a series of statement about a document in a pre-established order, and encoding it in standard format. And because RDF makes statements about a resource, the library community has so far been using it as an alternative to MARC coding; OCLC, for instance, enables us to translate MARC records into RDF Dublin Core and *vice versa*.

RDF, however, was conceived as something more than a means of describing electronic documents in a way analogous to our descriptions of physical documents. Because anything can theoretically be expressed as a unique identifier, URIs can be assigned to more than "documents" in the conventional sense. Individuals can be identified and described in much the same way. Even abstract concepts such as "love," "happiness," "work" and "desire" can be expressed within a particular community by adopting, by common consensus, a specific identifier. Nor is this an unheard-of step: after all, "sh 85078519" is the Library of Congress Control Number for

the Library of Congress Subject Heading, "Love," and "n 78087607" is the LCCN for the individual we know as Charles Dickens.

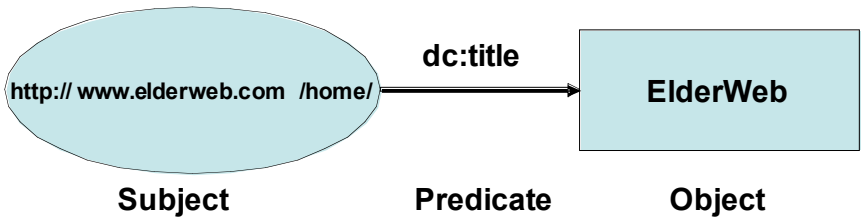

Figure 2. RDF statement

If RDF can describe intangible entities just as it can describe tangible documents, then it could provide a means for reference staff to identify and encode that most elusive of all things in libraries: the answer to a question. And just as they used to collect ephemeral physical documents that they thought would answer reference questions they were sure to encounter, so too might they use RDF to identify fragments of reliable Web documents that could answer a specific, frequently-posed question.

6. AAQ: Answers to Alzheimer's Questions

In order to understand more clearly how this would work, we designed a pilot project involving Websites related to Alzheimer's Disease. The pilot study proceeded in the following stages:

1. We envisioned an RDF framework which identified four related entities:
   - A user, who, among other attributes, possesses the attribute of knowledge level regarding Alzheimer's Disease: medical professional, layperson, etc.
   - A question posed by the user, related to Alzheimer's Disease;
   - An answer to that question;
   - A resource in which the answer can be found. (See Figure 3).

2. Using various websites developed by national and local Alzheimer's societies, as well as information from the National Library of Medicine and Health Canada, we developed a set of frequently-asked questions, classified according to various types of user: those who suffer from Alzheimer's disease, family caregivers, professional caregivers, and medical professionals.

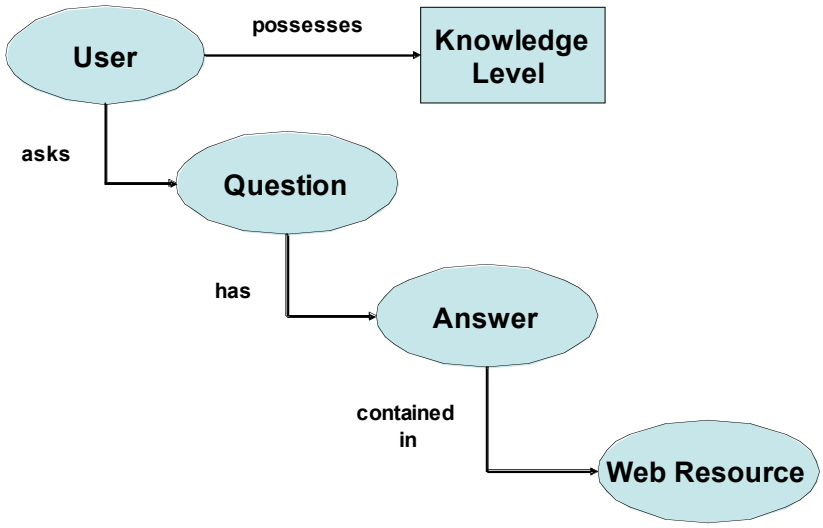

Figure 3. Initial entities

3. We located the answers to those questions in various Web resources which created by reputable and reliable sources.

4. Using the fragments of the websites that answered these specific questions, we simplified our scheme to two primary entities—a resource and an answer—and developed a metadata set, drawn from the Dublin Core, supplemented by our own additions (see Figure 4).

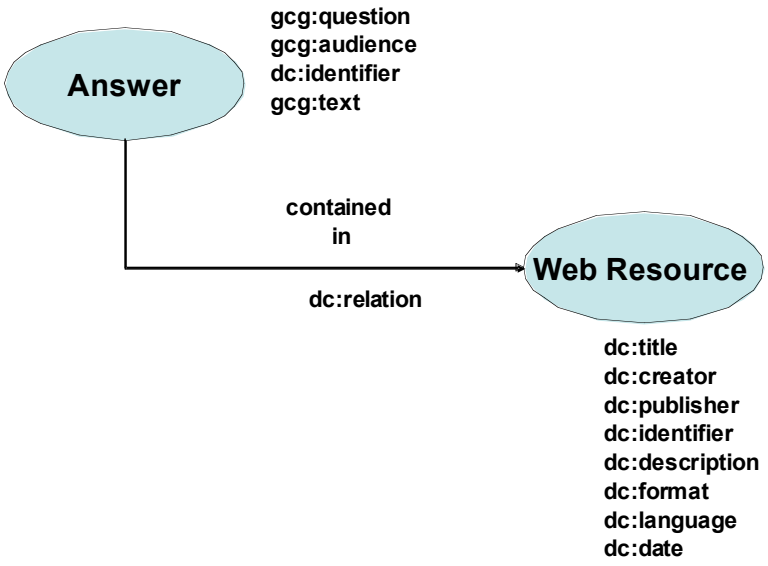

Figure 4. Working entities and attributes

For our resource guide, the Web resource had a series of Dublin Core attributes: title, creator, publisher, identifier, description, format, language and date. The Answer resource had the Dublin Core attributes of relation and identifier, as well as three attributes which we created in our own namespace, identified by the preface "gcg": question, audience and text.

5. We created RDF descriptions of both the answers to the questions and the resources from which the answers were drawn. A typical answer, for instance, might look like this:

```
<rdf:Description rdf:about="gm0004">
        <gcg:question>How is Alzheimer's disease treated?</gcg:question
        <gcg:audience>non-professional</gcg:audience>
        <gcg:text>Several medications are now available to treat some Alzheimer symptoms.
        These drugs are not a cure for the disease. They do not stop its progression. Ask your
        doctor if there is a treatment suitable for you.</gcg:text>
        <dc:relation>http://www.alzheimer.ca/english/haveAD/learning.htm</dc:relation>
</rdf:Description>
```

The website from which it was taken might be described like this:

```
<rdf:Description rdf:about="http://www.alzheimer.ca/english/haveAD/learning.htm">
        <dc:title>I Have Alzheimer's Disease: Learning About the Disease</dc:title>
        <dc:creator>Alzheimer Society of Canada</dc:creator>
        <dc:publisher>Alzheimer Society of Canada</dc:publisher>
        <dc:identifier>http://www.alzheimer.ca/english/haveAD/learning.htm
        </dc:identifier>
        <dc:description>frequently asked questions</dc:description>
        <dc:format>html/text</dc:format>
        <dc:relation>http://www.alzheimer.ca/english/haveAD/intro.htm</dc:relation>
        <dc:language>en</dc:language>
        <dc:date>2005/10</dc:date>
</rdf:Description>
```

5. We designed a prototype information portal which will enable users to use this RDF data to find the answers to these questions (see Figure 5).

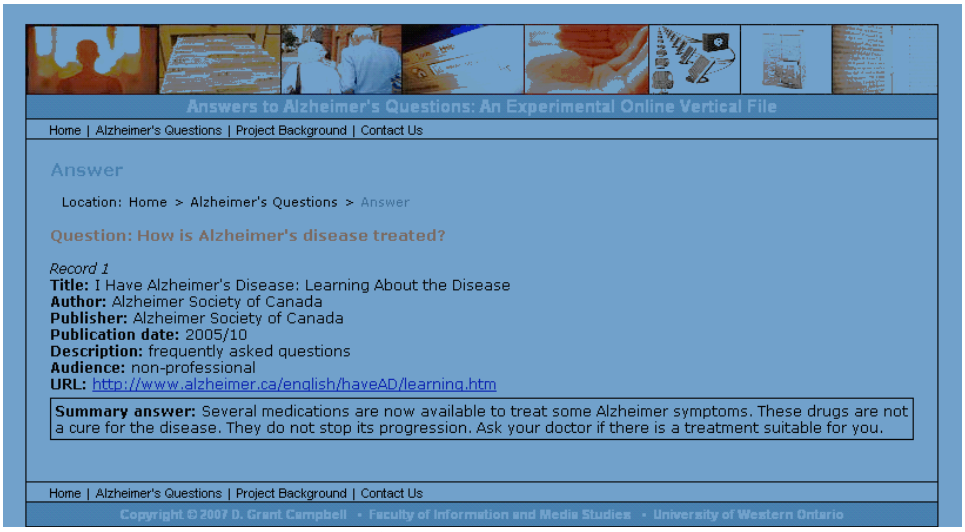

Figure 5: User Portal

7. Implications

Although much remains to be done before the portal is ready for testing, a number of implications have already become clear. First, the success of such a tool would depend on the development of an efficient and easy-to-use third-party annotation system for quickly creating the descriptions, so that librarians could maintain and add to the RDF database from numerous workstations, with very little effort.

Second, while the online system could be designed specifically for local users, some connection with other libraries and library systems could be facilitated through the use of the

common Dublin core namespace, permitting the aggregation of this RDF data into larger resources, while keeping the task of upkeep and maintenance distributed throughout the system.

In addition, the success of tools like this would require a redefinition of traditional library functions, rather than their wholesale rejection. The skills that originally led to such resources as clippings files and other methods of capturing ephemera in libraries are admirably suited to Semantic Web technologies and standards. But using such skills would shift resource description away from centralized technical services departments, into a closer relationship with other library services.

8. Conclusions

The prototype briefly described here rests upon the recognition of a close and bounded affinity between the sophisticated structure of traditional information organization and the standards of the Semantic Web, as opposed to the looser standards made possible by search engines, the Web 2.0 and tagging systems. While the Semantic Web has not attained the wholesale adoption originally predicted for it, its commitment to rigorous information structure has made it ideal for the migration of library systems and library techniques into Web environments. And an alliance between the Semantic Web community and the library community might serve to reinforce an assumption that there are still information environments that require rigorous and conscientious information organization, linked to the practices of information mediation and evaluation.

Such an alliance, however, will introduce fresh complexities into our distinctions between the physical document and the intellectual unit, as we move into an information environment where abstract entities such as questions and answers are described with the same language as information "resources" in the traditional sense. Classic knowledge organization will survive by letting go of its original commitment to the monograph level, using a combination of pragmatic skills acquired by cataloguers, subject analysts, collection development librarians and reference librarians to provide access to portions of documents that answer important questions.

Acknowledgements

Funding for this study has been provided by the Social Sciences and Humanities Research Council of Canada

References

Bray, T. (2003). The RDF.net challenge. *Ongoing* : retrieved May 28, 2007 from http://www.tbray.org/ongoing/When/200x/2003/05/21/RDFNet

Colburn, J.L. (1997). Horizontal wires replace the vertical files. *Medical reference services quarterly*, 16(3): 19-25.

Coyle, K., & Hillman, D. (2007). Resource Description and Access (RDA): Cataloging rules for the 20th century. *D-Lib Magazine*, 13(12): retrieved May 28, 2007 from http://dlib.org/dlib/january07/coyle/01coyle.html

Fast, K.V., & Campbell, D.G. (2005). 'I still like Google': University student perceptions of searching OPACs and the Web. *Proceedings of the American Society for Information Science and Technology*, 41(1): 138-146.

Lancaster, W. (1986). *Vocabulary control for information retrieval*. 2nd ed. Arlington: Information Resources Press.

Neuhaus, C. (1997). Developing a hypertext World Wide Web vertical file. *Collection building*, 16(1): 66-72.

Passin, T. (2004). *Explorer's guide to the Semantic Web*. Greenwich: Manning.

Shirky, C. (2003). The Semantic Web, syllogism and world view. *Clay Shirky's writings about the Internet*: retrieved May 28, 2007 from http://www.shirky.com/writings/semantic_syllogism.html

Shirky, C. (2005). Ontology is overrated: Categories, links and tags. *Clay Shirky's writings about the Internet*: retrieved May 28, 2007 from http://www.shirky.com/writings/ontology_overrated.html

Thomsen, E. (1997). The virtual vertical file. *Collection building*, 16(1): 30-35.