The Semantic Web and the Role of Information Systems Research[*]

A Position Paper[+] for the NSF-OntoWeb Invitational Workshop on
DB-IS Research for Semantic Web and Enterprises
http://lsdis.cs.uga.edu/SemNSF/
April 3-5, 2002

Frank Manola
The MITRE Corporation
202 Burlington Road, Bedford MA 01730
fmanola@mitre.org

1. Introduction

The Semantic Web is an exciting evolution of the current World Wide Web, and there is much activity surrounding it. However, in spite of the fact that the Web is clearly a large-scale information system, much of the activity taking place is in areas somewhat separated from the traditional database and information systems communities. This workshop has been organized to explore relationships between the Semantic Web (and related enterprise systems) and contributions that could come from the database and information systems communities.

Since this is a position paper, I'm going to use it to (briefly) state some "positions" which I hope will contribute to the Workshop discussions. These "positions" fall into three parts. First, I'm going to describe what I think the Semantic Web is, and isn't, since there is the potential for a certain amount of misunderstanding about that. Second, I will mention some research themes that I think are particularly interesting and important. Here, I will try to be particularly brief. This is because there are lots of obvious relationships between the Semantic Web and research areas in database and information systems technology. Many of these relationships have been pointed out in the Workshop Introduction and Background statement, I agree with all of them, I suspect that most of the Workshop participants do too, and I assume the details of these relationships will be discussed at the Workshop. Third, I will describe some things I think the database and information systems communities could do to become more involved with the evolution of the Semantic Web.

2. What the Semantic Web Is -- and Isn't

Recently, there has been an increasing amount published about the Semantic Web. Much of this has emphasized the potential of what might be accomplished using improved semantic descriptions of Web resources, based on ontologies and logic-based processing. However, in order to fully understand the dimensions of the Semantic Web, it's helpful to consider a few statements about it.

> "The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications." [W3C01]

> "The Semantic Web is a web of data, in some ways like a global database." and "Leaving aside the artificial intelligence problem of training machines to behave like people, the Semantic Web approach instead develops languages for expressing information in a machine processable form." [TBLmap98]

[+] This paper is available at the Workshop Web site as http://lsdis.cs.uga.edu/SemNSF/Manola-Position.doc, and is published in the Workshop Proceedings at http://lsdis.cs.uga.edu/SemNSF/SemWeb-DBIS-Workshop-Proc.pdf

> "The semantic web data model is very directly connected with the model of relational databases ... Indeed, one of the main driving forces for the Semantic web has always been the expression, on the Web, of the vast amount of relational database information in a way that can be processed by machines."
> [TBLnot98]

What is the significance of these statements? First, that a significant focus of the Semantic Web involves simply making Web data machine-processable. Of course, if programs are to process this data correctly, they must understand its semantics. However, *the Semantic Web vision doesn't really care how the programs acquire this understanding*. For example, this understanding may exist because the data uses a vocabulary (and associated semantics) that has been agreed on within a community, and the programs are written to use that vocabulary, just as with most databases now.

This isn't to say that having those semantics explicitly described in metadata (such as ontologies), using richer languages capable of associating more "meaning" with the data, won't be of tremendous importance. Obviously such explicitly-specified semantic information will be increasingly important in the Web environment, given its scale, number of data sources, and their autonomy and heterogeneity. For example, such information will be crucial in such areas as integrating and making sense of information from heterogeneous sources, optimizations of various kinds, advanced types of concurrency control, and so on. However, lots of very effective Semantic Web applications (and certainly lots of near-term ones) will not necessarily involve extensive use of machine-processable semantic information. As someone said on the xml-dev email group, "Semantic Web my behind! I just want to order a pizza, not have mozzarella explained to me!" It will be a tremendous job to migrate the current Web to the point where all of the data is machine-processable even without machine-processable semantics, and where all of the current tools we have available in database systems are available. Hence, it will be important not to neglect technologies in areas other than those directly involving machine-processable semantic information, and to integrate those technologies more thoroughly into the Web environment.

Another significant thing about these statements is that the Semantic Web is considered as being very much like a database (this, of course, is something we've suspected all along!), and in particular, like a relational database. Semantic Web ideas are currently based on what, from a database perspective, can be thought of as a restricted relational model. In this model, URIs (Uniform Resource Identifiers [RFC 2396]) identify the things people want to talk about (termed "resources"). URIs can be thought of as a superset of the URLs (Uniform Resource Locators) used in web browsers. As with URLs, different persons or organizations can independently create URIs, and use them to identify things. However, unlike URLs, URIs are not limited to identifying things that have network locations, or use other computer access mechanisms. In fact, a URI can be created to refer to anything anyone wants to talk about, including

- network-accessible things, such as an electronic document, an image, a service (e.g., "today's weather report for Los Angeles"), or a collection of other resources.
- things that are *not* network-accessible, such as human beings, corporations, and bound books in a library.
- abstract concepts that don't physically exist, like the concept of a "creator" property

Given URIs to identify resources, people then write property/value pairs describing those resources in the form of triples of **subject** (resource), **property**, and **object** (resource or literal value). That's the basic idea behind RDF [RDF02], and languages layered on it such as DAML+OIL [DAML01]. The basics of this type of model have had a long history in the database community, in the form of binary relations, since the triples can also be written in the form **property (subject, object)** (anyone remember binary relational models?). However, the RDF / Semantic Web variant involves an important extension, in that it is grounded in the Web: the subjects, objects, *and properties* in RDF statements are URIs, and hence have global identity. Moreover, in many designs these URIs can be dereferenced (you can use them as URLs and point a browser at them) to retrieve Web data that describes the resources. The result is an interesting mixture of relational and object-oriented ideas (since the URIs are effectively a form of object identifier), and of data and metadata. Moreover, compared with XML, RDF really does play the role of the "relational

model of the Web", since, for example, it requires that relationships not be recorded implicitly in data structures (such as XML nested elements), but instead must be made explicit.

Finally, I think it's important to emphasize that this is the *Web* we're talking about: the largest distributed system in existence, and the largest distributed database in existence. It's accessed by, *and contributed to by*, thousands and thousands of individuals and companies, and it implicitly contains all database data (as well as all the services) that can be accessed through it. This represents a grand challenge for information systems technology if I ever saw one! Even without advanced semantics processing, the Semantic Web is an evolution of this existing Web that will provide a database in which programs can do useful things whose scope is hard to imagine. Even though this evolution will happen to different extents, at different rates, in different parts of the Web, over time, it not only will happen, it is happening right now. So, in my view, when the Workshop Introduction and Background statement describes the Semantic Web as "one of the recent unifying visions", it seriously understates the case, as does calling it a "vision" at all. "An emerging reality" would be more accurate.

3. Research Themes

As I said in the Introduction, there are lots of obvious relationships between the Semantic Web and research areas in database and information systems technology. Many of these relationships have been pointed out in the Workshop Introduction and Background statement; I agree with all of them, and I suspect that most of the Workshop participants do too. Advances in any of these areas would most likely be relevant and useful in building the Semantic Web. The important thing to remember in addressing these relationships is to focus on the particular aspects of the Web that differ from older application environments for these technologies, particularly:

- The scale and openness of the Web, in particular, the difficulty of creating closed-world models, either for instances or for definitions, and the existence of multiple, autonomous sources and authorities.
- The need to deal with potentially rapid and continuous change, both in instances and in definitions, and the need for the "system" to continue to operate in spite of these changes.

In addition, there are a few other themes I want to mention. First, lots of discussions of the Web in database circles have been in terms of how database technology can be adapted to the Web. This is all very well, as long as the basic focus is not "how do we shoehorn database ideas into the Web", but rather “how do we manage data on the Web?". That is, start from the top down, as if this were a new problem, and then see whether database ideas, or new ones, are more appropriate. We can still talk (to a large extent) about the same things, since we know that database technology has a lot of background and potentially relevant technology to bring to the party, and there are interesting adaptations of the technology that may be necessary. The focus, however, needs to be on leveraging our experience in managing large amounts of data, not on database technologies per se (as if we think the Web is a nail, because we have a hammer). Also, this approach may make it clearer to other communities that we’re not just talking about applying this technology in the context of DBMSs that are somehow attached to the Web, but actually focusing on managing the Web data (with DBMSs playing a role, together with browsers, servers, proxies, etc., as part of a larger architecture). Then we can start talking about Semantic Web topics using techniques like: making Web data more structured, adding schemas and semantics, searches using first structured, and then “smart”, data, transactions, managing URIs as identifiers (there’s a vast OODBMS literature dealing with identity, for example), conceptual schema languages, integration of database and knowledge base technologies, etc.

As a related issue, there should be some focus on “data management architectures” (not “DBMS architectures”). For example, Web search engines don't try to do query optimization by treating the Web sites they intend to search as part of a distributed database. Instead, they pull that data into indexes and associated data to create giant caches (i.e., additional architectural components) that can be more readily searched as "real databases". We need to be thinking about how similar architectural extensions of the Web and associated systems can be applied to handle other “database-like” problems in the different environment of the Web.

Regarding capturing "semantics" (or conceptual design), people typically have "ontologies" in mind when they design databases. We've been training them to think like this for years. However, we haven't had a clean way (or much reason) to capture all that information in the richer models that knowledge representations provide (without using a separate tool often not linked to the database). Now we have better reasons to push this harder. The information systems community needs to *adopt* ontologies, make them a more explicit part of their data management strategies, and then investigate how to use them in database processing. All the prior work in conceptual languages and deductive databases is relevant here. At the same time, we need to make sure to point to the need to support "industrial scale" management of large amounts of definitions and content that continually change (leveraging, e.g., prior work on versioning and schema modification), and the whole design and implementation process.

It goes without saying (but I'll say it anyway) that our prior work on integration of heterogeneous data is relevant, and needs to be pushed and extended. A major effort is required to make the process of doing this integration easier. It can't all be automated (e.g., writing "articulation axioms"), or the requirements known ahead of time, so we need to investigate how to make it easier for people to help, both in at design time (as before) and *at run time* (we've got instant messaging between people; how about between an agent or query engine and a person, along the lines of "I don't understand exactly what 'sale price' consists of; can you clarify?"). Such an intervention might be quite reasonable in the context of some types of Web applications.

Work on integration also needs to consider how to use the Web (and its technologies) as a potential organizing basis and environment or infrastructure for information sharing (i.e., use of the Web as a *tool* for solving data management problems, not just as the object of data management technologies). A key observation here seems to me that part of solving the data integration problem is providing a data integration environment that enables all the pieces to be put together (and, for example, enables people to have access to and potentially reuse metadata and schema designs). The Web is a widely-available shared environment which can make this happen (the development of shared repositories for XML DTDs and schemas is an example of this idea in action).

Finally, a lot more work needs to be done on the various notions of "contexts" [Guha91, MB98, TBLcg01]. This involves not only recording and using (e.g., for mediation) metadata about the definitions and assumptions used by different data collections. It also involves being able to integrate information described using languages of different descriptive capabilities, as well as being able to integrate different "reasoning engines" that operate on that information, on the same Web, and have them interoperate. The Web allows people with different reasoning capabilities to interoperate (up to a point!), and this needs to be a goal for our systems as well. The Introduction and Background statement for the Workshop mentions the "tradeoff between expressiveness and computability". We need to be prepared for the fact that people will make different decisions about this tradeoff, and try to cope with all of them.

4. How Do We Make It Happen?

One of the themes of the Workshop is that there has not been much focus on database and information systems topics that appear to be relevant to the Semantic Web, and that research in these areas has not been pursued as part of Semantic Web development activities. Here are a couple of suggestions on what to do about that.

4.1 Get involved in the relevant communities

There's an old political proverb that says, "you can't beat something with nothing". Another way of saying this is "you have to get in the game in order to win it (or even score)". It's all very well to point out connections between our research and research we claim is needed on the Semantic Web in manifestos and other documents circulated among the DB-IS community. But we're (apparently) complaining that in spite of these connections, either (a) we're not getting funding, or (b) no one's paying attention, or (c) all of the above. All this does no good if most of the action is taking place somewhere else. And the fact is the main Semantic Web action is taking place in Web-oriented forums, not in DB-IS ones.

Even if all we want is to encourage research funding (and who isn't interested in that?), no one would expect to just build a DBMS and expect eager customers to seek them out in order to buy it. Instead, you'd expect to have to send people out to talk to potential customers and tell them (a) about your products and (b) how they can meet those customers' needs. Part of what the DB-IS community needs to do more of is get into Web-oriented forums with the message of what their technologies have to offer. There's a good deal of (very important) participation from the database community in activities like W3C's XML Query activity, for example, but that was an easy connection to make! Where else in Web-oriented forums is the relevance of our technology being demonstrated (or even suggested)? Lacking that demonstrated relevance, the Web community will solve what they see as their problems themselves. And they aren't necessarily going to start reading SIGMOD Record or attending PODS conferences. They need to have this information brought to their attention more explicitly (and often!).

One obvious approach is to see if there aren't relevant W3C activities to become involved in. In many cases, you don't need to be a member to participate in relevant W3C activities (see http://www.w3.org/Consortium/#public). You also don't need a big travel budget (most formal W3C activity takes place via email and teleconferences, with relatively few face-to-face meetings). The W3C is not a conventional standards activity; it's more a technology incubation activity. If you want to incubate technology relevant to the Semantic Web, look into getting involved, directly or indirectly.

Another approach is to establish a presence in public email lists where Semantic-Web-related topics are discussed, e.g.:

- xml-dev (http://www.xml.org/xml/xmldev.shtml)
- www-rdf-interest (http://www.w3.org/RDF/Interest/#discussion)
- rdf-logic (http://lists.w3.org/Archives/Public/www-rdf-logic/)

Similarly, publish in the increasing number of conferences related to the Semantic Web. The point is to get in a position where DB-IS-related ideas can be injected into the relevant discussions.

4.2 Make the Relevance of DB-IS Ideas Concrete

There's another (not-so-old) proverb that says, "the Internet is based on rough consensus and running code". The point here is that once you're talking to the right people, *how* do you talk to them? Running code is a very convincing way to talk to people in the Web community about how DB-IS technology can solve problems they are interested in. For example, it strikes me that many current DB-IS projects could do a better job of more directly indicating the relationship of IS-DB work to the Web (for an example of work that does this, look at Stefan Decker and Sergey Melnick's work at Stanford; http://www-db.stanford.edu/).

This doesn't mean more work on DBMSs for XML (as important as that may be), published in SIGMOD. That's "data management for XML data", not "Web data management". Web sites in the DB-IS community are just as accessible (or ought to be) as anyone else's. Put your technology up on a web site, let people use it from there or download it, and use those Web-oriented forums to tell the Web community about it. In many cases, this wouldn't add much, if any, additional overhead to the way the projects operate now, but would add enormously more impact as far as the Web community is concerned.

References

[Guha91] R. V. Guha, "Contexts: A Formalization and Some Applications", Ph.D. Dissertation, Stanford University, 1991, http://www-formal.stanford.edu/guha/index.html

[MB98] " John McCarthy and Sasa Buvac, "Formalizing Context (Expanded Notes)", 1998, http://www-formal.stanford.edu/jmc/mccarthy-buvac-98/index.html

[RDF02] RDF Core Working Group page, http://www.w3.org/2001/sw/RDFCore/

[RFC 2396] RFC 2396 - Uniform Resource Identifiers (URI): Generic Syntax , August 1998 http://www.isi.edu/in-notes/rfc2396.txt

[TBLmap98] Tim Berners-Lee, "Semantic Web road map", September 1998, http://www.w3.org/DesignIssues/Semantic.html

[TBLnot98] Tim Berners-Lee, "What the Semantic Web can represent" [sometimes known as "What the Semantic Web is not"], September 1998, http://www.w3.org/DesignIssues/RDFnot.html

[TBLcg01] Tim Berners-Lee, "Conceptual Graphs and the Semantic Web", 2001, http://www.w3.org/DesignIssues/CG.html

[W3C01] W3C (World Wide Web Consortium) Semantic Web Activity page, http://www.w3.org/2001/sw/