What's new about the Semantic Web? Some questions

Karen Sparck Jones
Computer Laboratory, University of Cambridge
William Gates Building, JJ Thomson Avenue, Cambridge CB3 0FD, UK
sparckjones@cl.cam.ac.uk

Preamble

It is not completely clear what the Semantic Web (SW) is meant to be. Statements about it differ or, rather, the interpretations of statements about it differ. The grand view is that the SW is the core model of the way the world is, expressed in a manner that supports reasoning about this world. The modest view is that the SW is the minimal apparatus of shared generic terminology that can be used to send some carrier pigeon messages from one universe of discourse to another. The 'authorised' view presented in Berners-Lee, Hendler and Lassila (2001) might be regarded as intermediate. It treats the key notion of ontology as a structure of well-defined, i.e. unambiguous, concepts standing for objects, properties, relations etc, which has an accompanying logic allowing inference about the concepts. This is much like the terminological and assertional components familiar from AI knowledge representation. There is indeed no claim that everything that users may want to talk about (or all information) can be captured in this way, either descriptively or inferentially, But the belief is that, even with restricted conceptual coverage and limited inference, it will be possible to throw a large semantic net over the Web, that does far more to catch the information fish than purely syntactic hooks and with much less effort for the human fisherman. This large net can, conveniently, be made by combining many little nets with a few well-made loops. The whole rests upon a proper formal language syntactic apparatus (XML, RDF).

Unfortunately, there are serious problems about the core SW idea of combining substantive formal description with world-wide reach, i.e. having your cake and eating it, even if the cake is only envisaged as more like a modest sponge cake than the rich fruit cake that AI would like to have. Specifically, definitions for general concepts, necessarily expressed in natural language, are hostages to fortune, and there is no guarantee that they can link domains consistently. Thus arguing that one can achieve seamless semantic connectivity over the Web is in fact relying on a far stronger version of the SW than the notion of bridging, with its aura of engineering efficiency, suggests.

The talk examined the implications of the SW idea, starting from the grander version and moving towards the modest one, drawing in particular on the experience of the library and information world. The lessons to be drawn from this are that there are early limits

to what you can do with semantic metadata, and that the proper arena for information access is natural language.

SW: the 'high end' version

The strong SW assumption is that for *real* information access, i.e. for natural language information processing, we need much more than surface word matching. We need a universal world model, and a formal (or controlled) language for description and reasoning with respect to the model. This is not a new idea: Leibniz and Wilkins were promulgating it in the 17th Century. But they could not carry it through. The grand notion of a Universal Character to talk of the world, and an encyclopedia about the world, was eventually realised only in Roget's Thesaurus, a specifically linguistic object.

This may look like a mountain bringing forth a mouse. It must be possible to do better. In particular, if IT is the source of the challenge the Web represents as a holder of information, IT must also, somehow, make the crucial difference to the form, content, or use of a modern Universal Character.

IT does make a difference. The World Wide Web is a world wide success. We have mighty processor power, untold data bytes, endless connectivity. The Web is already routinely, autonomously, supporting humans in their information chores, like filing. It is increasingly helping humans in their information actions, like retrieving. Hopefully, with the SW, it will start replacing humans in their information decisions. Proper information management embeds decisions, it doesn't just enable them. Doing this needs logical reasoning on knowledge, not just statistical operations on words.

There is an underlying assumption here. This is that one world model is sufficient for multiple tasks, and so is one reasoning apparatus, just as they are for humans. In fact, information access is also being taken as a generic task. But are text retrieval, document filtering, question asnwering, information extraction, passage selection, summarising, ... data query, ... all just variations of the same process, in the sense of having common world modelling and reasoning requirements?

It might look as if they are. For example, suppose we have a document that talks about making gardens, notes that pansies are useful and come in many colours. Then if we want to retrieve documents on flower gardens, we can exploit a model relationship between pansy and flower to get the document. The model and subsumptive inference improve matching. Again, if we ask a question about what pansies are good for, we can exploit the the fact that pansies have colours and are used (as flowers) for nice things, together with the fact that gardens are nice things, to reply that pansies are good for nice gardens. This is a more demanding form of information access, but not different in kind. Further, viewing summarising as a form of access, we can imagine some inferences focused on the document's remarks about pansy colours and pansy utility to state that pansy colours suit gardens. In all of this we operate on a combination of prior and current document-derived model content.

The tasks differ, but the examples nevertheless suggest that they can all rely on one ontology and logic, the same, necessary, apparatus of objects, properties, roles and implication relations.

Here we need to look at what happened in the library and retrieval world, i.e. in the main information access world in the past; and we should also look more deeply at the tasks.

Lessons from information retrieval

First (and very simplistically), in cataloguing and indexing, there were global classification schemes, like the Universal Decimal Classification. These were for books and had assignment to a single class as default. Indexing was as specialised, downward, as the classification allowed, while searching could generalise upward. The advantage of the approach was that inference was very easy. The problems were that there were word/world ambiguities in the classification model and class uncertainties in searching.

To overcome these problems, especially with the growth of the specialised journal literature, document description schemes were developed that used concept labels and role relations, defined by facets or operators. With these, descriptions could be constructed for papers, based on the structural decomposition of complex notions to component elements linked by general relations. To achieve matching in searching, these descriptions could be simplified by, say, weakening relationships. This strategy had the advantage of more descriptive flexibility, but problems with domain biases in description and viewpoint selectivity when searching, so the schemes were not general.

The need to make indexing less restrictive led to the now-conventional approach using subject heading systems or thesauri, with Boolean connectives to combine headings for descriptions. Indexing was based on a relatively undemanding notion of topic labelling, and searching on limited structural (broad/narrow/related term) substitution. The advantages are fair flexibility and generality, but the Boolean connectives define very limited links between description components.

Modern retrieval schemes move further away from control, using (key)words or strings with simple coordination. Papers are viewed more as texts than as bodies of content for indexing, while searching is simply best match by term count or sum of weights. Indexing is based on extracted not assigned keys, with free variation in combination for matching. The advantages are high flexibility and great generality; the disadvantage is that inference is submerged.

Overall, the trend in document characterisation has been away from lexical normalisation and towards relational simplification, i.e. towards decreasing ontological expressiveness, decreasing epistemological commitment, and decreasing inferential power. But this has been correlated with wider application and better task performance.

This cannot be an accident, and retrieval research has shown that it is not an accident. But this does not necessarily imply that the SW approach is irrelevant. Document retrieval is a minimal form of information access, and primarily text focused, because it presumes that the user has, for the best reasons, to read the text themself. Do other tasks involve more demanding forms of access that do require a world model and a logic?

Other tasks

If we consider question answering and take e.g. a report on the ladies final match at Wimbledon as the source for answers, we can obtain many different answers to the

same question, without any one being obviously 'the' right one. If we take information extraction, e.g. from a report on a type of camera, we can pull out many nuggets, with no definite limit on which. With summarising, say of a stock market report, we can produce a range of quite different, but equally plausible, summaries.

It is clear that such information access, or management, tasks are very complex in how they make input-output connections, and very variable in how they determine output preferences. This seems to imply that world modelling and logical inference are the correct base for automation, though there may also be significant use of surface language information, for example in extracting subject-verb-object triples as factoids, or in selecting statements providing answers.

But thus supposing we continue to believe that we do need model and inference, there are lessons to be learnt from work on natural language access to databases. Databases have formal models and database search is logical. But there are a lot of hidden semantics in the guise of user knowledge, as work on informal natural language access (in the 1980s) made very clear. A natural language interface needs a domain model as well as the abstract data model, both to bridge the gap between the user's approach and the database reality, and to trap the user's inadvertent excursions outside the bounds of the database. Even simple conversions from relationships between things to attributes of things can require a surprising amount of apparatus: a domain model can be much more substantial than the data model that is required to characterise the database itself. But such a domain model will still be very application-particular, where the SW explicitly requires generality.

SW: the 'middle end' version

These two lines of experience, from document retrieval and database query, suggest a need to modify the original strong version of the SW. The modified version of the SW is that the SW ontology is an 'upper model'. Everyone shares this common upper model, which enables useful basic or initial access to more detailed information, and more specific application ontologies can be attached below it.

But there are some lessons here from AI work on common sense knowledge and reasoning. It was widely accepted that specialised knowledge rested on a substrate of 'ordinary' knowledge and that this would be shared across applications. Thus for example the CS knowledge base could include ordinary knowledge about liquids, e.g. that they flow, and a specialised model e.g. of the blood system, would be 'appended' to the CS model. But in doing this it is necessary to handle the information required to answer questions about blood clotting correctly. Experience showed that general CS models get very complicated and that connecting specific models to them is very tricky (as was apparent with CYC).

Lessons from natural language processing

All of this suggests that the world modelling approach is too exigent to be feasible for anything on the SW scale. There are, moreover, alternative lessons in NLP practice, e.g. from the NIST/ARDA programme on question answering from text. The data here is unrestricted news material, in vast quantity, where we cannot envisage deep world modelling or any conventional database strategy. Many of those who are tackling this task get

help from WordNet, a broad-cover, general-purpose, concept organisation that is explicitly linguistic. It captures linguistic relations and structures, e.g. synonymy, verb frames, so world structure is only implicit; and it indicates the habitual and generic, not the mandatory and particular, about word use. But though WordNet is about words, not things, and offers only weak, essentially subsumptive, inference, it can usefully support operations on text to derive answers to questions from it.

However experience with question answering also raises issues about the nature and extension of information access tasks. For example, answering questions by extracting answers from bulk text data shows how ill-defined the notion of correct answer is, even when apparently obvious. There may be answers with additional material, answers which do not give but may suggest an answer, etc. The user has always to bring their own knowledge and context to interpreting the system output: the system cannot have enough of either to do the whole job, so too tight an approach to what a question answering system should do is misconceived.

Such observations imply that it is wrong to go for any depth in the SW idea. We should recognise that shallow text operations - select, match, show - are right for information access. Information is primarily conveyed by natural language and this has to be shown to the user for them to assess. The SW's role should thus not be that of the total intermediary, but as a little light underpinning, primarily format oriented, with very minimal semantic tagging.

To illustrate this, we can take the Columbia University Newsblaster system that offers automatic summaries of news stories. The role of SW-type shallow metadata in such a system could be to identify news stories where these are mixed with other material, perhaps also category tags like dates, places, agencies in story headers. But all the substantive summarising is by shallow natural language text processing, combining statistical and syntactic techniques. Perfectly usable summaries can be produced without any attempt to characterise text content in a knowledge representation language.

SW: the 'low end' version

Confining the SW to field tagging is essentially high-level cataloguing of the familiar library or museum kind, exemplied by 'author', 'title', 'publication date' and so forth. Done properly, this is far from trivial, as the substantial *Anglo-American cataloguing rules* demonstrates. For example, is the author exactly what appears on the title page or some specific person? But though proper cataloguing is not for amateurs, it is not necessary for useful cataloguing to go overboard on rules. It is more sensible to go for a practical approach to field tagging, even if it is not perfect, and to focus on what can be done with the text in the fields, even simply using statistical processing strategies and, perhaps, lightweight NLP. This will be rough information access, because the field tagging will not always be right and the text processing will not be comprehensive, and it will rely on the user. But it will be a win for the user because they will get something and, in particular, something amplified by the normally available, as well as necessary, discourse context.

The rationale for this approach is well illustrated by museum or sales catalogues, that refer to very heterogenous and complex non-text objects. A typical sales catalogue has an

entry with a picture, a header e.g. "A Charles II parcel-gilt cagework cup, circa 1670", and a body e.g. "The cylindrical silver gilt body crested with pierced and chased sleeve of scrolling flowers and foliage incorporating various birds ...". It is not too difficult to get broad field types from the header e.g. 'object type: cup'. But even though the descriptive body uses specialised terms of art e.g. "sleeve", "scrolling", it could not be decomposed into an unambiguous formal equivalent. Access to the content of the body is therefore rightly through the natural language words used themselves.

Postscript

A discussion contributor reiterated the statement that the SW is only intended as a bridging device, with no more semantic power than that needed to provide relatively uncontroversial and hence common ontological hooks into more specialised domains, and with ontology concepts defined on pages anyone can read. In attacking the deep version of the SW I was attacking a strawman. My argument is not against the practicality and utility of specialised ontologies, or against the possibility of linking specialised subdomain ontologies. My argument is that the ontology approach to the everything that the Web *as a whole* deals with, once it goes beyond some very high-level (albeit still semantic) tags like 'image', 'book', perhaps 'news', which can apply, and be reasonably similarly understood, across the Web, is a misconceived enterprise if it assumes that there is something better than natural language as a *general* means of expressing, and hence accessing, information. Indeed the SW notion of ontology definition pages in natural language is tacitly acknowledging the early limits to formal rigour. It is much better to bite the natural language bullet right away.

References

T. Berners-Lee, J. Hendler and O. Lassilia. 'The Semantic Web', *Scientific American*, May 2001.

M. Gorman and P.W. Winkler (Eds.) *Anglo-American cataloguing rules*, 2nd ed, 1988 revision, Chicago, ILL: American Library Association, 1988.

http://newsblaster.cs.columbia.edu