



What is the Semantic Web and what will it do for eScience?

Yorick Wilks

Research Associate
Oxford Internet Institute
www.dcs.shef.ac.uk/~yorick

Abstract

The paper discusses what kind of entity the proposed Semantic Web (SW) is, in terms of the relationship of natural language structure to knowledge representation (KR). It argues that there are three distinct views on the issue: first, that the SW is basically a renaming of the traditional AI knowledge representation task, with all the problems and challenges of that task.

If that is the case, as many believe, then there is no particular reason to expect progress in this new form of presentation, as all the traditional problems of logic and representation reappear and it will be no more successful outside the narrow scientific domains where KR seems to work even though the formal ontology movement has brought some benefits. The paper contains some discussion of the relationship of current SW doctrine to representation issues covered by traditional AI, and also discusses issues of how far SW proposals are able to deal with difficult relationships in parts of concrete science.

Secondly, there is a view that the SW will be the WorldWideWeb with its constituent documents annotated so as to yield their content or meaning structure more directly. This view of the SW makes natural language processing central as the procedural bridge from texts to KR, usually via a form of automated Information Extraction. This view is discussed in some detail and it is argued that this is in fact the only way of justifying the structures used as KR for the SW. There is a third view, possibly Berners-Lee's own, that the SW is about trusted databases as the foundation of a system of web processes and services, but it is argued that this ignores the whole history of the web as a textual system, and gives no better guarantee of agreed meanings for terms than the other two approaches.

Introduction

This paper is concerned with the issue of what kind of object the Semantic Web is to be and, in particular, to ask about its semantics in the context of the relationship between knowledge representations and natural language itself, a relationship concerning which this paper wishes to express a view which will appear below. This is a vast, and possibly ill-formed issue, but the Semantic Web is no longer simply an aspiration in a magazine article (2001) but a serious research subject on both sides of the Atlantic, with its own conferences and journal. So, even though it may not exist in a demonstrable form, in the way the WWW itself plainly does exist, it is a topic for research and about which fundamental questions can be asked, as to its representations, their meanings and their groundings, if any.

The position adopted here is that the concept of the Semantic Web (SW) has two distinct origins, and this persists now in two differing lines of SW research: one, closely allied to notions of documents and natural language (NL) and one not. These differences of emphasis or content in the SW carry with them quite different commitments on what it is to interpret a knowledge representation and what the method of interpretation has to do with meaning in natural language.

We shall attempt to explore both these strands here, but our assumptions will be with the NL branch of the bifurcation above, a view that assumes that natural language is, in some clear sense, our primary method of conveying meaning and that other methods of conveying meaning (formalisms, science, mathematics, codes etc.) are parasitic upon it. This is not a novel view: it was once associated firmly with the philosophy of Wittgenstein (1953), who we shall claim is slightly more relevant to these issues than Hirst's immortal, and satirical, line that 'The solution to any problem in AI may be found in the writings of Wittgenstein, though the details of the implementation are sometimes rather sketchy' (2000).

The Semantic Web and AI/NLP

The Hirst quotation above serves to show that any relation between philosophies of meaning, such as Wittgenstein's, and classic AI (or GOFAI as it is often known: Good OldFashioned AI) is not an easy one. GOFAI remains committed to some form of logical representation for the expression of meanings and inferences, even if not the standard forms of the predicate calculus; most issues of the AI Journal consist of papers of this genre.

Many have taken the initial presentation (2001) of the SW (by Berners-Lee, Hendler and Lassila) to be just a restatement of the GOFAI agenda in new and fashionable WWW terms: they describe a system of services, such as fixing up a doctor's

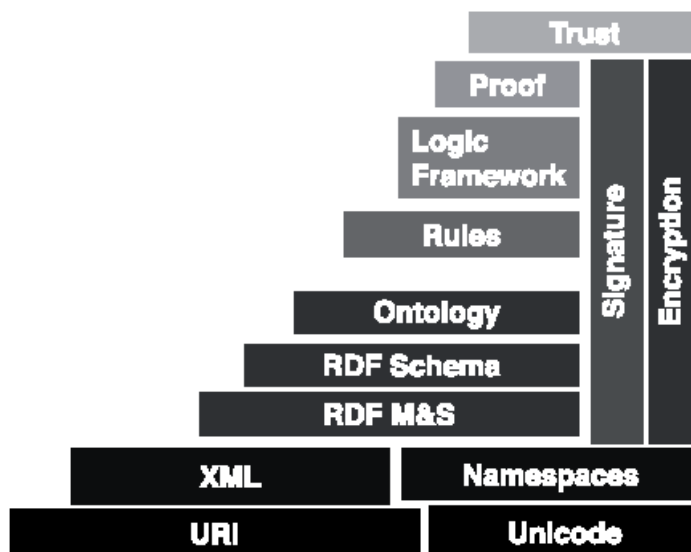
appointment for an elderly relative, which would require planning and access to both the databases of the doctor's and relative's diaries and synchronizing them. This kind of planning behaviour was at the heart of GOFAI, and there has also been a direct transition (quite outside the discussion of the SW) from decades of work on formal knowledge representation in AI to the modern discussion of ontologies. This is clearest in work on formal ontologies representing the content of science (e.g. Horrocks, 2005), and many of the same individuals (e.g. Pat Hayes) have transferred discussion and research from one paradigm to the other. All this has been done within what one could call the standard KR assumption within AI, and one that goes back to the earliest work on systematic KR by McCarthy and Hayes (1969), that the predicates in such representations merely look like English words but are in fact formal objects, loosely related to the corresponding English, and without its ambiguity, vagueness and ability to acquire new senses with use. We shall return to this assumption below. These assumptions of GOFAI have certainly been apparent in both the original SW paper and some of what has flowed from it, but there have been at least two other traditions of input to what we now call work on the SW and I shall discuss one in some detail: namely, the way in which the SW concept has grown from the traditions of document annotation.

Before leaving GOFAI, it must be noted that very little of the complex theories of knowledge representation appears within SW contributions so far: from McCarthy and Hayes fluents (McCarthy and Hayes, 1969), McCarthy's later autoepistemic logic (1990), Hayes' Naïve Physics (1979), Bobrow and Winograd's KRL (1977), to name but a few prominent examples. Some continuity of goals between GOFAI and the SW has not meant continuity of tradition and this is both a gain and a loss: a gain because of a simpler schemes of representations which are probably computable; a loss because of the lack of sophistication in current schemes of the DAML/OIL (<http://www.w3.org/TR/daml+oil-reference>) family, and the problem of whether they now have the representational power for the complexity of the world, whether common sense or scientific, a point we shall return to later.

Annotation and the lower end of the Semantic Web

If one looks at the classic SW diagram from the original *Scientific American* paper (see Figure 1), the tendency is always to look at the upper levels: rules, logic framework and proof, and it is these that have caused both critics and admirers of the SW to say that it is the GOFAI project by another name. But if one looks at the lower levels one finds Namespaces and XML, which are all the products of what we may broadly call NLP (natural language processing) obtained by annotations of texts by a range of NLP technologies we may conveniently call IE (information extraction).

**Figure 1. Levels of annotation and objects in the Semantic Web
(from Berners-Lee et al., 2001)**



It is useful to remember that available information for science, business and everyday life, still exists overwhelmingly as text; 85% of business data still exists as unstructured data (i.e. text). So, too, of course does the WorldWideWeb, though the proportion of it that is text is almost certainly falling. And how can the WWW become the SW except by information being extracted from natural text and stored in some other form; the standard candidates provided by IE (e.g. Cunningham et al., 1997) being either a database of facts extracted from text or annotations on text items, stored as metadata either with or separate from the texts themselves. XML, the annotation standard which has fragmented into a range of cognates for non-text domains (e.g. TimeML, VoiceML etc.) is the latest in a history of annotation languages that attach codings to individual text items so as to indicate information about them, or what should be done with them in some process, such as printing. The annotation languages grew from origins in publishing metadata (the Stanford roff languages, then, above all, Knuth's Tex, later LaTeX), as well as semi-independently in the humanities community for formalizing the process of scholarly annotation of text. The Text Encoding Initiative (TEI) adopted SGML, a development of Goldfarb's (1997, publication date) GML, which then became the origin of HTML (as a subset), then XML and the annotation movement in NLP that initially underpinned, as its first phase, IE technology. There were early divisions over exactly how and where the annotation of text was to be stored for computational purposes; particularly between SGML, on the one hand, where annotations were infixed into the text with additional characters (as in LaTeX), which made the text harder to read and, on the other hand, the DARPA research community that produced a functioning IE technology which tended to prefer storage of annotations (indexed by spans of characters in the text) separately as metadata, a tradition preserved in, for example, the GATE language processing platform from Sheffield (Cunningham et al., 1997), which now underpins many of the SW projects in Europe (e.g. Boncheva et al., 2003; Ciravegna et al., 2003).

IE is now a technology with some twenty-five years of history: it began with Leech's CLAWS4 program (Leech et al., 1994) to do automatic part-of-speech tagging in 1966: the first program systematically to add to a text 'what it meant' even at a low level. IE now reliably locates names in text, their semantic types, and relates them together by learned structures called templates into forms of fact, objects virtually identical to the RDF triple stores at the basis of the SW: not quite logic, but very like IE output. IE began with automating annotation but now has what we may call annotation engines based on machine learning (Brewster et al., 2001) which learn to annotate in any form and in any domain.

Extensions of this technology have led to effective question-answering systems against text corpora in well-controlled competitions and, more recently, the use of IE patterns to build ontologies directly from texts (Brewster et al., 2005). Ontologies can be thought of as conceptual knowledge structures, which organize facts derived from IE at a higher level. They are very close both to the traditional Knowledge Representation goal of AI, and they occupy the middle level in the original SW diagram. I shall return to ontologies later, but from now I only want to draw attention to the obvious fact that the SW rests on some technology with the scope of IE, probably IE itself, to annotate raw texts and derive first names, then semantic typings of entities, fact databases, and later ontologies.

This view of the SW, which is not the only one, as I emphasized at the beginning, is the one that underlies most work on the SW and webservices in Europe (Brewster et al., 2001). On this view, the SW could be seen as a conversion from the WWW of texts by means of an annotation process of increasing grasp and vision, one that projects notions of meaning up the classic SW diagram from the bottom. Richard Braithwaite (1956) once wrote an influential book on how scientific theories obtain the semantic interpretation of 'high level' abstract entities (like neutrinos or bosons) from low level data; he named the process one of *semantic ascent* up a hierarchically ordered scientific theory. The view of the SW under discussion here, which sees NLP and IE as its foundational processes, bears a striking resemblance to that view of scientific theories in general.

Blurring the text and program distinction

The view of the SW sketched above has been that the IE technologies at its base, technology that adds 'the meaning of a text' to the text in varying degrees and forms, causes a blurring of the distinction between language and knowledge representation because the annotations are themselves forms of language, sometimes very close indeed to language they annotate. I shall return to this point below, but here I want to note in parallel that this process at the same time blurs the distinction between programs and language itself. This is of relevance to our theme here in that it has never been totally accepted that knowledge is always and essentially in declarative form.

This distinction has already been blurred historically from both directions:

1. Texts are really programs (one form of GOFAI)
2. Programs are really texts

As to the first, there is Hewitt's (1972) contribution to an NLP book, devoted to how to plan the building of a wall and containing the claim that 'language is essentially a side effect' of programming and knowledge manipulation. Longuet-Higgins (1975) also devoted a paper to the claim that English was essentially a high-level programming language. Dijkstra's view of natural language was in essence that it was really not up to the job it had to do, and would be better replaced by precise programs, which is almost a form of the first view. On the other side, a smaller group, is what one might term the Wittgensteinian opposition, and I will cite my own version (2005), which is the view that natural language is and always must be the primary knowledge representation device, and all other representations, no matter what their purported precision, are in fact parasitic upon language—in the sense that they could not exist if NL did not—and they can never be wholly divorced from NL, in terms of their interpretation and use. This paper is intended as a modest contribution to that tradition: but a great deal more can be found in a dialogue with Nirenburg in Nirenburg and Wilks (2001).

Systematic annotations are just the most recent bridge from language to programs and logic, and it is important to remember that not so long ago it was perfectly acceptable to assume that a knowledge representation must be derivable from an unstructured form, i.e. natural language. Thus Woods in 1975:

'A KR language must unambiguously represent any interpretation of a sentence (logical adequacy), have a method for translating from natural language to that representation, and must be usable for reasoning.'

The emphasis there is on a method of going from one to the other, that is, from the less to the more formal, a process which inevitably imposes a relation of dependency between the two. This gap has opened and closed in different research periods: in the original McCarthy and Hayes (1969) writings on KR in AI, it is clear that NL was thought vague and dispensable. The annotation movement associated with the SW can be seen as closing the gap in the way in which we have discussed it, but it is quite clear that the first view stated in this paper (roughly GOFAI) still holds to the old McCarthy–Hayes position on the issue.

The separation of the annotations in metadata (versus leaving them within the text as in LaTeX style annotations) has strengthened the possibility of the dispensability of the original language from which the representation was derived, whereas the infixing of annotations in a text suggests the whole (original plus annotations) still forms some kind of linguistic object. Notice here that the 'dispensability of the text' view is not dependent on the type of representation derived—in particular to strongly logical representations. Schank (1972) certainly considered the text dispensable after his Conceptual Dependency representations had been derived, because he believed them to contain all the meaning of the text, implicit and explicit. This is the key issue that divides opinion here: how can we know that any representation whatsoever contains all and only the meaning content of a text? What could it be like to know that?

The standard philosophical problems may or may not just vanish as we push ahead with annotations to bridge the gap from text to meaning representations, whether or not we then throw away the original text. David Lewis in his 1970s (1972) critique of Fodor and Katz, and of any non-formal semantics, would have castigated all annotations as ‘markerese’: his name for marking up language with markers that are still within NL and thus not reaching to any meaning outside language. The Semantic Web movement, as described in this section of the paper at least, takes this criticism head on and continues, hoping URIs and « popping out of the virtual world » (e.g. by giving the web representation your—real world—phone number!) will solve semantic problems. That is to say, it accepts that the SW is based on language via annotations and that will provide sufficient ‘inferential traction’ with which to run web-services. But is this plausible? Can all you want to know be put in RDF triples, and can this then support the reasoning required? But agents so-based do seem to work in practice. In the end, of course, nothing will satisfy a critic like Lewis except a web based on a firm (i.e. formal and extra-symbolic) semantics and effectively unrelated to language at all. But a century of experience with computational logic should by now have shown us that this cannot be had outside narrow and complete domains, and the SW may be the best way of showing that a non-formal semantics can work effectively, just as language itself does, and in the same ways.

An Information Retrieval (IR) critique of the semantics of the SW

Sparck Jones (2004) in a critique of the SW has returned to a theme she has deployed before against much non-empirically based NLP, such as ontology building; in her phrase ‘words stand for themselves’ and not for anything else, and that claim has been the basis of successful IR search in the WWW and elsewhere. Content, for her, cannot be recoded in a general way especially if it is general content, as opposed to some very specific domain, such as medicine, where she believes ontologies may be possible. As she puts it: IR has gained from ‘decreasing ontological expressiveness’.

Her position is a restatement of the traditional problem of ‘recoding content’ by means of other words (or symbols closely related to words, such as thesauri, semantic categories, features, primitives etc.), as annotation attempts to do this on an industrial scale. Sparck Jones’ key example is (in part): ‘A Charles II parcel-gilt cagework cup, circa 1670’. What, she asks, can be recoded there beyond the relatively trivial: {object type: CUP}?

What, she argues, of the rest of that (perfectly real and useful) description of an artefact that cannot be rendered other than in the exact words of the catalogue?

This is a powerful argument, but the fact remains that content can be expressed in other words: it is what dictionaries, translations and summaries routinely do. Where she is right is that GOFAL researchers are quite wrong to ignore the continuity of their predicates and classifiers with the language words they resemble (an issue

discussed at length in Nirenburg and Wilks, 2001). What can be done to ameliorate this impasse?

One method is that of empirical ontology construction from corpora (REFS), now a well-established technology, even if not yet capable of creating complete ontologies. This is a version of the Woods quote above by which a KR representation (an ontological one in this case) must be linked to some natural language text to be justifiably derived: the derivation process itself can then be considered to give meaning to the conceptual classifier terms, in a way that just writing them down a priori does not. An analogy here would be with grammars: when linguists wrote these down 'out of their heads' they were never much use as input to programs to parse language into structures. Now that grammar rules can be effectively derived from corpora, parsers can, in turn, produce structures from sentences using such rules.

A second method is based on the observation that we must take 'words as they stand' (Sparck Jones) but perhaps not all words are equal; perhaps some are aristocrats, not democrats. Perhaps what were traditionally called 'semantic primitives' are just words but also special words: forming a special language of translation or coding, albeit one that is not pure but ambiguous, like all language.

If there are such 'privileged' words, perhaps we can have explanations, innateness (even definitions) on top of an empiricism of use. It has been known since Olney et al. (1968) that counts over the words used in definitions in actual dictionaries show a very clear set of primitives on which most definitions rest.

By the term 'empiricism of use', I mean the approach that has been standard in NLP since the work of Jelinek (Jelinek and Lafferty, 1991) in the late 80s and which has effectively driven GOFAI approaches based on logic out of NLP. It will be remembered that Jelinek attempted at IBM a machine translation system based entirely on machine learning from bilingual corpora. He was not ultimately successful but he changed the direction of the whole NLP field as researchers tried to reconstruct by empirical methods the linguistic objects on which NLP had traditionally rested: lexicons, grammars etc. The barrier to further advances in NLP by these methods seems to be the 'data sparsity' problem to which Jelinek originally drew attention, namely that language is 'a system of rare events' and a complete model, at say the trigram level, for a language seems impossibly difficult to derive, and so much of any new corpus will always remain uncovered by such a model.

The Web as corpus and the hope of much larger language models

I want to suggest here that it may now be possible, using the whole web, to produce much larger models of a language and to come far closer to the full language model that will be needed for tasks like complete annotation and automatically generated ontologies. These results are only suggestive and not complete yet, but they do seem to make the data for a language much less sparse and without loss by means

of skipgrams. The Wittgensteinian will always want to look for the use rather than the meaning, and nowhere has more use than the whole web itself.

It has been noted already (Kilgarriff and Grefenstette, 2001) that the web itself can now become a language corpus in principle, even though that corpus is far larger than any human could read in a lifetime, as a basis for language learning. A rough computation shows that it would take about 60,000 years of constant reading for a person to read all the English documents on the WWW at the time of writing. But the issue here is not a psychological model and this need not deter us: Moore (2004) has noted that current speech learning methods would entail that a baby could only learn to speak in a hundred years of exposure to data, but this has been no drawback to speech technology—in the absence of anything better—provided the method is effective. Its effectiveness has been shown by experiments by e.g. Grefenstette (2003) who has shown that the most web-frequent translation of a word pair formed from all possible translation equivalent pairs in combination is invariably also the correct translation.

What follows is a very brief description of the kind of results coming from the REVEAL project (Guthrie et al., 2006), which take a 1.5 billion word corpus from the web and ask how much of a test corpus is covered by the trigrams of that large training corpus, both as regular trigrams and as skipgrams which are trigrams consisting of any discontinuity of items with a maximum window of four skips between any of the members of a trigram. The 1.5 billion word training corpus gives a 67%+ coverage by trigrams of 1000 word test texts in English (Figure 2).

Since the whole web is hard to get at, could we go a simpler way such as skipgrams? Suppose, as a way of extending the training corpus, we consider skipgrams, and take:

Chelsea celebrate Premiership success.

The trigrams will be:

Chelsea celebrate Premiership
celebrate Premiership success

But one-skip trigrams will be:

Chelsea celebrate success
Chelsea Premiership success

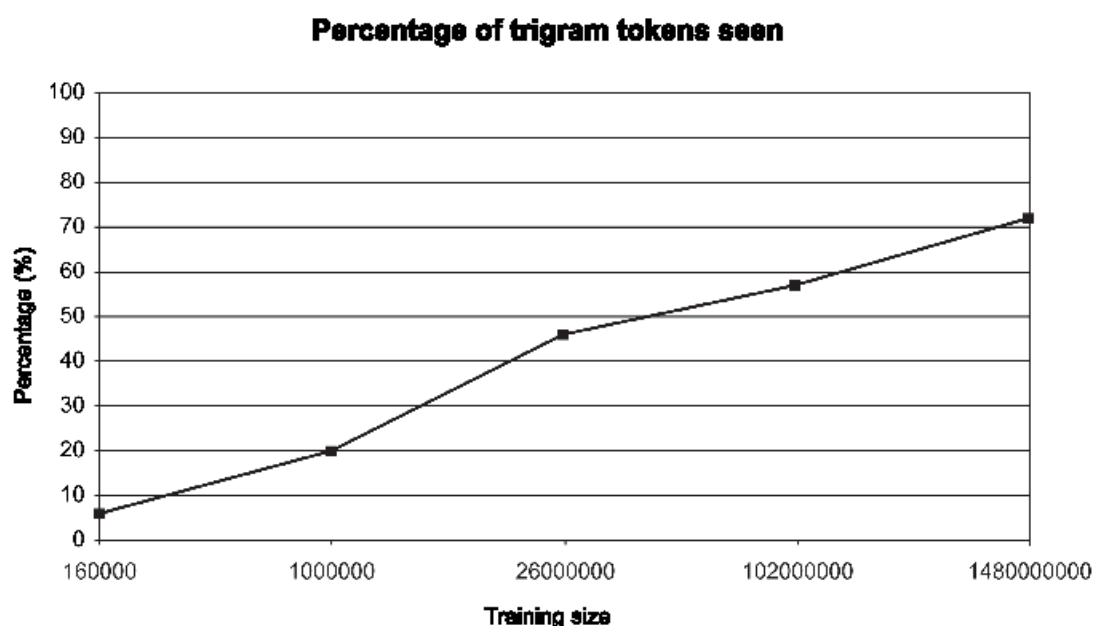
Which seem at least as informative, intuitively, and our experiments suggest that, surprisingly, skipgrams do not buy coverage at the expense of producing nonsense. Indeed, recent work shows data sparsity for training may not be quite as bad as we thought: using skipgrams can be more effective than increasing the corpus size! In the case of a 50 million word corpus, similar results are achieved using skipgrams as by quadrupling corpus size.

This illustrates a possible use of skipgrams to expand contextual information to get something much closer to 100% coverage with a (skip) trigram model, thus combining greater coverage with little degradation, this achieving something much closer to Jelinek's original goal for an empirical corpus linguistics.

Thus, we obtained 74% coverage with four-skip trigrams. This suggests by extrapolation that it would need 75×10^{10} words to give 100% trigram coverage. Our corpus giving 74% was 15×10^8 words, and Greffentette (2003) calculated there were over 10^{11} words of English on the web in 2003 (i.e. about 12 times what Google indexes), so the corpus needed for complete coverage would be about seven times the full English web in 2003, which is presumably somewhat closer to today's English web.

All this is preliminary and tentative, but it suggests an empiricism of usage may be stronger (with bigger corpora) than Jelinek thought at the time he wrote.

Figure 2. Percentage of trigrams seen with training corpus size



These corpora are so vast they cannot offer a model of how humans process semantics, so a cognitive semantics based on such usage remains an open question. However, one possible way forward would be to adapt skipgrams so as to make them more likely (perhaps with the aid of a largescale fast surface parser) to pick up Agent-Action-Object triples in very large numbers. This is an old dream going back to Wilks (1965) where they were seen as Wittgensteinian 'forms of fact', later revived by Greffentette as a 'massive lexicon' and now beginning to be available as inventories of surface facts at ISI (Hovy, 2005). The point of relevance here is that these will not be very different from RDF triples, and might offer a way to massive content on the cheap, even simpler than that now offered by machine learning based

IE. If anything were possible along these lines, then NLP would be able to provide the base semantics of the SW more effectively than it does now, by making use of some very large portion of the WWW as its corpus.

The SW and point-of-view phenomena

One aspect of the SW that relates directly to issues of language and language processing is that of a SW as promoting or filtering a point-of-view on the Internet: so that, for example, it might be possible to use the annotations to prevent me seeing any web pages incompatible with The Koran, for example, and that might be an Internet-for-me that I could choose to have. However, there is no reason why such an annotated web should, as some have argued (e.g. Nelson, 2005) necessarily impose a single point of view. The technology of annotations is quite able to record two quite separate sets of annotation data (as meta-data) for the same texts, and no uniformity of point of view is either necessary or desirable.

It is worth remembering that the underlying page-rank technology of Google (Page et al., 1998) is itself a point-of-view phenomenon, not in the sense of controlling consistency, but as promoting in rank what is most believed. This is the basis of the criticism many make of Internet search in general, arguing that what is most believed is not necessarily what is true, as when, for example, most people are said to have believed the Earth was flat, even though scientists believed it was round and we may take it as true that it was round at the time. However, it is by no means clear that this is a good perspective on knowledge in general, particularly in view of an increasing number of phenomena where human aggregates seem better able to predict events than experts, a subject that has been of much interest recently (Surowiecki, 2004) to economists, and which has a clear relation to ontologies and other information structures built not from authority but from amalgamations of mass input, sometimes described as the 'wiki' movement or 'folksonomies' (e.g. <http://www.flickr.com/>).

The point-of-view and web-for-me issues are of wider importance than speculating about religious or pornographic censorship: they are important because of the notion of their being a correct view of the world, one that the SW and its associated ontologies can control in the sense of controlling the meanings of terms (the subject of this paper) as well as the wider issue of consistency with a received view of truth. The nearest thing we have to a received view of truth in the C21, in the Western world at least (and the restriction is important) is that of science, and not least because the web was developed by scientists and serves their purposes most clearly, even though they do not now control the Internet as they did at its inception, one could see the SW as an attempt to ensure closer links between the future web and scientific control. This emerges a little in what I have called the third view of the SW and its obvious similarity of spirit to Putnam's view that, in the end, scientists are the guardians of meaning.

I do not share this view but I do not dismiss it, partly because meaning and control of information would be safer in the hands of scientists than many other social groups.

However, because I believe that meanings cannot be constrained by any such methods, all such proposals are in the end hopeless.

But let us stay for a moment with a more rigorously scientific sociology of a SW and think what that might mean, remembering that there is more than one view of what scientific activity consists in: in Popper's view (1959), for example, it is the constant search for data refuting what is currently believed. An SW-for-me for a scientist following a Popperian lifestyle would therefore consist of a constant trawl for propositions inconsistent with those held at the core. This is, one might say almost the exact opposite of the normal information gathering style which is to accept attested information (by some criterion, such as page rank or authority) *unless contradicted*. In the philosophy of science, there is little support for Popper's views on this—a far more conventional view would be that of Kuhn (1982) that what he calls 'normal science' does not seek refutation or contradiction—and I mention this here only to point up that (a) scientific behaviour in information gathering is not necessarily a guide for life but that (b) such differences in approach to novel information can map naturally to different strategies for seeking and ranking incoming information with respect to a point-of-view or personal SW.

In conclusion, there is a quite different implementation of points-of-view that we may expect to see in the SW, but which will be more an application of NLP than a theoretical extension: the idea that some users will need an interface to any web, WWW or SW, that eases access by means of conversational interaction, which is itself a classic NLP application area. One can see the need from the Guardian correspondent (8.11.03) who complained that '... the internet is killing their trade because customers ... seem to prefer an electronic serf with limitless memory and no conversation.'

On this view (see Wilks, 2004) we will need personalized agents on our side: what we have called Companions. Their motivation, for the general population at least, is that the Internet is wonderful for academics and scientists, who invented it, but does not serve less socially competent citizens well, and certainly not excluded groups like the old. Recent evidence in Britain suggests that substantial chunks of the population are turning away from the Internet, having tried it once. As the engines and structures powering the Internet become more complex, it will also be harder for the average citizen to get what they could get from an increasingly complex structure, i.e. the Semantic Web, which will be used chiefly by artificial agents. Companion agents can be thought as persistent agents, associated with specific owners, to interface to the future Internet, to interact with it and in some ways to protect them from its torrent of information about them. They will know their owner's details and preferences in great detail, probably learned from long interaction, and will be needed even for simple present-day tasks such as search to avoid the repetitive and distressing cycles of disambiguation, error and correction we all suffer daily. They may also need them to organize the mass of information they will hold on themselves (e.g. the EPSRC Memories for Life project: <http://www.memoriesforlife.org/>), and they can only be built with NLP technology—specifically dialogue technology.

A third view of what the SW is

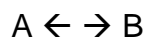
I have delayed a promised third view of the SW, different from both the GOFAI and NLP views that I have contrasted in this paper. That is, in my view, one close to Berners-Lee's own vision of the SW, as expressed in Berners-Lee et al. (2005), one that emphasizes databases as the core of the SW: namely databases, the meanings of whose features are kept constant and trustworthy by a cadre of guardians of their integrity, a matter quite separate from both logical representations (dear to GOFAI) and to any language-based methodology, of the kind described in this paper. Berners-Lee's view deserves careful discussion and consideration that cannot be given here, but it will have the difficulty of any view (like GOFAI) that seeks to preserve predicates, features, facets or whatever from the NLP vagaries of sense change and drift with time. We still 'dial numbers' when we phone but that no longer means the actions it did a few decades ago; so not even number-associated concepts are safe from time.

In some ways Berners-Lee's view has the virtues and defects of Putnam's later theory of meaning (Putnam, 1975/1985): one in which scientists became the guardians of meaning, since only they know the true chemical nature of, say, molybdenum and how it differs from the phenomenally similar aluminium. It was essential to his theory that the scientists did not allow the criteria of meaning to leak out to the general public, lest they became subject to change. Thus, for Putnam only scientists knew the distinguishing criteria for water and deuterium oxide (heavy water) which seem the same to most of the population but are not. Many observers, including this author, have argued this separation cannot be made, in principle or in practice.

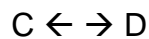
It should also be made clear that the first view of the SW, that it was an extension of the GOFAI project, does not entail that recent SW work has corporate memory of the sophisticated work done in AI during the 1970s and 1980s; on the contrary, as this topic is too large for this paper, it seems almost certain that, in its quest for computability, logic-driven SW work has opted for simplistic representations that are simply not able to represent the complexity of scientific knowledge that the SW will require.

The SW and the representation of tractable scientific knowledge

Kazic (2006) poses many issues close in spirit to those of this paper, but against a background of expert knowledge of biology that is hard to capture here without more exposition than was needed in Biocomputing Proceedings, where she published. Broadly, and using arbitrary names for terms like ‘thymidine phosphorylase’, there are two symmetric reactions of ‘cleavage’ we may write as:



and



An enzyme Z (actually EC 2.4.2.4) catalyzes both according to the standard knowledge structures in the field (KEGGs maps: <http://www.genome.ad.jp/kegg/kegg1.html>). But Z is not in the class Y (a purine nucleoside) and so should not, in standard theory, be able to catalyze the two reactions above, *but it does*. There is a comment in the KEGG maps saying that Z can catalyze reactions like those of another enzyme Z' (EC 2.4.2.6) under some circumstances, where Z' is a Y, but its reactions are quite different from Z and they cannot be substituted for each other, and neither can be rewritten as the other. Moreover, Z has apparently contradictory properties, being both a statin (which stops growth) and a growth factor. Kazic asks ‘so how can the same enzyme stimulate the growth of one cell and inhibit the growth of another?’ (p. 2).

We should leave this poor attempt to state the biological facts in this non-specialist form, but simply note that something very odd is going on here, something that Marxists might once have hailed as a dialectical or contradictory relationship. It is certainly an abstract structure that challenges conventional knowledge representations and is far more complex than the standard form of default reasoning in AI that if anything is an elephant it has four legs but nonetheless Clyde, undoubtedly an elephant, has only three. The flavour of the phenomena here is that of extreme context dependence, that is to say, that an entity behaves quite differently, indeed in opposite fashions, in the presence of certain other entities than its general type would suggest. Languages are, of course, full of such phenomena, as when ‘cleave to the Lord’ and ‘cleave a joint’ mean quite opposite things, and we have structures in language representation for describing just such phenomena, though there is no reason at the moment to believe they are of any assistance here.

The point Kazic is making is that it will be a requirement on any SW that represents biological information and licences correct inferences that it can deal with phenomena as complex as this, which at first sight seem beyond those normally thought of as within a standard ontology dependent on context free relations of inclusion and the other standard relations we expect: ‘To ensure the scientific validity of the Semantic Web’s computations, it must sufficiently capture and use the

semantics of the domain's data and computations' (p. 2). In connection with the initial translation into RDF, for example, she continues:

'Building a tree of phrases to emulate binding ... forces one to say explicitly something one may not know (e.g. whether the binding is random or sequential, what the order of any sequential binding is ...). By expanding the detail to accommodate the phrasal structure, essential and useful ambiguities have been lost.' (ibid.)

The last quotation is most revealing about the structure of science, the degree to which in parts it remains a craft skill, even in the most advanced and technical modern areas. If that were not the case, being forced to be more explicit and to remove ambiguities could only be a positive influence. The quotation brings out clearly the dilemma in some parts of advanced science that intend to make use of the SW: that of whether the science is explicit enough and well understood enough to be formally coded, a question quite separate from the issues of whether the proposed codings (from RDF to DAML/OIL) have the representational power to express what is to be made explicit. If this is the case, then Biology is not so different from ordinary life as we may have thought, certainly not so different from the language of auction house catalogues, in Sparck Jones' example, where the semantics remains implicit, in the sense of resting on our human interpretation of the words of annotations or comments (in this case in KEGG maps).

The analogy here is not precise, of course; we have made the point that the representational styles in the current SW effort have, in some degree sacrificed representational sophistication to computational tractability (as, in another way, the WWW itself did in the early 90s). It may well be that when some of the greater representational powers in GOFAL work are brought to bear, the KEGG-style comments may be translated from English phrases, with an implicit semantics, to the explicit semantics of ontologies and rules. It is what we must all hope for. In the case of Sparck Jones' C16 cup, the problem does not lie in any knowledge representation issue, but only in the fact that the terms involved are all so precise and specific that no generalizations exist—no auction ontology, for example—that would enable the original English to be thrown away. One could always translate these words into another language or an explicit numbering of all the concepts in it, but there is no representational saving to be made there in either case (see here, as always, McDermott, 1981).

Kazic goes on to argue that an effect of these difficulties of explicitness is that 'most of the semantics are pushed onto the applications', where the web agents may work or not, but there is insufficient explicitness to know why in either case. This is a traditional situation: as when a major AI objection from AI to the connectionist/neural net movement was that, whether it worked or not, nothing was served scientifically if what it did was not understood. Again, in so far as she is right, and there is not enough SW data yet to be sure, it is completely against the spirit of the SW that its operations should be unnecessarily opaque or covert; and that becomes even clearer if one sees the SW as the WWW 'plus the meanings' where only additional, rather than less, explicit information would be expected.

Discussions in this general area normally resile themselves from the discussion of more traditional ontological enquiry, namely *what things are there in the world*, so as to get to where the arguments currently go on, or should go on. Ancient questions have a habit of returning to bite one at the end, even though, in this paper, we have taken as a robust position, in the spirit of Quine (op. cit.) that whatever we put into our representations—concepts, sets, etc.—have existence, at least as a polite convention, so we can continue with the discussion. But it may be the case that a fully explicit SW has to make ontological commitments of this more traditional sort, at least as regards the URIs the points where the SW meets the world of real things. But it may be interesting to note that our initial scientific examples of genes are by no means as straightforward as we pretended, and that this will have impact on the view we have styled (after Putnam, 1970) ‘Scientists as guardians of meaning’, and one we argued had relations to Berners-Lee’s own view of SW scientists as trusted keepers of data-bases that give the SW integrity of meaning. Putnam, it will be remembered, argued that only scientists know the real scientific criteria that distinguished water from heavy water, and they should keep these to some extent away from the public and everyday usage, so that the meanings of these terms did not change, with disastrous consequences for science.

Suppose we ask again, what are the ontological ‘objects’ in genetics, say in the classic *Drosophila* data base FlyBase (Morgan et al., 2003)? FlyBase ultimately grounds its gene identifiers—the formal gene names—in the sequenced *Drosophila* genome and associates nucleotide sequences parsed into introns, exons, regulatory regions etc. with gene ids. However, these sequences often need modifying on the basis of new discoveries in the literature (e.g. new regulatory regions ‘upstream’ from the gene sequence are quite frequently identified, as understanding of how genes get expressed in various biological processes increases). Thus the ‘referent’ of the gene id. changes and with it information about the role of the ‘gene’. However, for most biologists the ‘gene’ is still the organizing concept around which knowledge is clustered so they will continue to say the gene ‘rutabaga’ does so-and-so quite happily even if they are aware that the referent of rutabaga has changed several times and in significant ways over the last decade. The curators and biologists are, for the most part, happy with this, though the argument that the *Drosophila* community has been cavalier with gene naming has been made from within it. This situation, assuming this non-expert description is broadly correct, is of interest here because it shows there are still ontological issues in the original sense of that word: i.e. as to what there actually IS in the world. More precisely, it calls into question Putnam’s optimistic theory (1970, cited elsewhere in this paper) that meaning can ultimately be grounded in science, because, according to him, only scientists know the true criteria for selecting the referents of terms. The *Drosophila* case shows this is not so, and in some cases the geneticists have no more than a hunch, sometimes false in practice, that there are lower level objects unambiguously corresponding to a gene id., in the way that an elementary molecular structure, say, corresponds to an element name from Mendeleev’s table.

Conclusion

The main argument of the paper has been that contemporary NLP offers a way of looking at usage in detail and in quantity—even if the huge quantities required now show we cannot easily relate them to an underlying theory of human learning and understanding. We can see glimmerings, in machine learning studies, of something like Wittgenstein's 'language games' (1953) in action, and of the role of key concepts in the representation of a whole language. Part of this will be some automated recapitulation of the role primitive concepts play in the organization of (human-built) ontologies, thesauri, and wordnets.

I would argue that NLP will continue to underlie the SW in several different ways: but chiefly it is the way up to a defensible notion of meaning at conceptual levels (in the original SW diagram) based on lower-level empirical computations over usage. The paper's aim is definitely not to claim logic-bad, NLP-good in any simple minded way, but to argue that the SW will be a fascinating interaction of these two methodologies, unlike the WWW which has been basically a field for NLP. It also assumes that, whatever the limitation on current SW representational power we have drawn attention to here, the SW will continue to grow in a distributed manner so as to serve the needs of scientists, even if it is not perfect. The WWW has already shown how an imperfect artefact can become indispensable.

Acknowledgements: This paper is indebted to many discussions with colleagues within the AKT project (Aktive Knowledge Technologies: EPSRC Interdisciplinary Research Centre, 2001-6), as well as with Ted Nelson, Arthur Thomas and Christopher Brewster. The passage on *Drosophila* owes a great deal to conversations with Ted Briscoe.

References

- Berners-Lee, T., Hendler, J. and Lassila, O. (2001) *Scientific American*. May 35-43.
- Berners-Lee, T. (2005) Keynote paper in BCS Workshop on the Science of the Web, London.
- Bobrow, D. and Winograd, T. (1977) An overview of KRL, a knowledge representation language. *Cognitive Science* 1:3-46.
- Bontcheva, K. and Cunningham, H. (2003) Information Extraction as a Semantic Web Technology: Requirements and Promises. Adaptive Text Extraction and Mining workshop, 2003.
- Braithwaite, R. (1956) *Scientific Explanation* (Cambridge University Press: Cambridge).
- Brewster, C., Ciravegna, F. and Wilks, Y. (2001) Knowledge Acquisition for Knowledge Management. Position Paper in Proceedings of the IJCAI-2001 Workshop on Ontology Learning held in conjunction with the 17th International Conference on Artificial Intelligence (IJCAI-01), Seattle, August, 2001.
- Brewster, C., Iria, J., Ciravegna, F. and Wilks, Y. (2005) The Ontology: Chimaera or Pegasus. Proc. Dagstuhl Seminar on Machine Learning for the Semantic Web, 13-18 February 2005.
- Ciravegna, F. and Wilks, Y. (2003) Designing adaptive information extraction for the Semantic Web in Amilcare. In: S.Handschuh and S.Staab (eds) *Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications* (IOS Press).
- Cunningham, H., Humphreys, K., Gaizauskas, R. and Wilks, Y. (1997) GATE—a TIPSTER-based General Architecture for Text Engineering. Proc. of the TIPSTER Text Program Phase III (Morgan Kaufmann: CA).
- Goldfarb, C.F. (1977) SGML: The Reason Why and the First Published Hint. *Journal of the American Society for Information Science* 48:656-661.
- Greffenstette, G. (1994) *Explorations in automatic thesaurus discovery* (Kluwer: Boston).
- Greffenstette, G. (2003) Using the Web as a language model. Lecture, University of Sheffield, 2003.
- Greffenstette, G. (2004) The scale of the multi-lingual Web. Talk delivered at Search Engine Meeting 2004, The Hague, The Netherlands, 19-20 April 2004.
- Guthrie, D., Allison, B., Liu, W., Guthrie, L. and Wilks, Y. (2006) A Closer Look at Skip-gram Modelling. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy.

- Hayes, P.J. (1979) The Naive Physics Manifest. In: D.Michie (ed.) Expert Systems in the Micro-Electronic Age (Edinburgh University Press: Edinburgh), pp. 242-270.
- Hewitt, C. (1973) Procedural Semantics. In: Rustin (ed.) Natural Language Processing (Algorithmics Press: New York), pp. 107-123.
- Hirst, G. (2000) Context as a spurious concept. Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February 2000, pp. 273-287.
- Horrocks, I. (2005) Description Logics in Ontology Applications. KI/Tableaux 2005, Koblenz, Germany, September 2005.
- Hovy, E. (2005) Key toward large-scale shallow semantics for higher-quality NLP. Proc. 12th PACLING Conference, Tokyo, Japan.
- Jelinek, F. and Lafferty, J. (1991) Computation of the Probability of Initial Substring Generation by Stochastic Context Free Grammars. Computational Linguistics 17:315-323.
- Kazic, T. (2006) Putting the semantics into the semantic web: how well can it capture biology? Proc. Pacific Symposium in Biocomputing 11:140-151.
- Kilgarriff, A. and Greffenstete, G. (eds) (2001) The Web as Corpus. Computational Linguistics (Special Issue).
- Kuhn, T.S. (1962) The Structure of Scientific Revolutions (University of Chicago Press: Chicago).
- Leech, G., Garside, R. and Bryant, M. (1994) CLAWS4: The tagging of the British National Corpus. Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) Kyoto, Japan, pp. 622-628.
- Lewis, D. (1972) General Semantics. In: D.Davidson and G.Harman (eds) The Semantics of Natural Language (Kluwer: Amsterdam).
- Longuet-Higgins, H. (1972) The Algorithmic Description of Natural Language. Proc. Roy. Soc. Lond. B 182:255-276.
- McCarthy, J. (1990) Formalizing common sense: papers by John McCarthy (Ablex: Norwood, NJ).
- McCarthy, J. and Hayes, P. (1969) Some Philosophical Problems from the Point of View of Artificial intelligence. Machine Intelligence 4 (Edinburgh University Press: Edinburgh), pp. 463-502.
- McDermott, D. (1981) Artificial Intelligence meets Natural Stupidity. In: J.Haugeland (ed.) Mind Design (Bradford: Montgomery, VT), pp. 143-160.
- Moore, R.K. (2003) A comparison of data requirements for ASR systems and human listeners. Proc. EUROSPEECH 2003, pp. 2581-2584.
- Morgan, A., Hirschmann, L., Yeh, A. and Colosimo, M. (2003) Gene Name Extraction Using FlyBase Resources. ACL Workshop on Language Processing in Biomedicine, Sapporo, Japan, pp. 18-26.
- Nirenburg, S. and Wilks, Y. (2001) What's in a symbol. Journal of Theoretical and Empirical AI (JETAI), pp. 9-23.

- Olney, J., Revard, C. and Ziff, P. (1968) Some monsters in Noah's Ark. Research memorandum, Systems Development Corp., Santa Monica, CA.
- Page, R., Brin, S., Motwain, R. and Winograd, T. (1998) The pagerank citation algorithm: bringing order to the web. In 7th WWW Conference, pp. 107-117.
- Popper, K.R. (1959) *The Logic of Scientific Discovery* (Routledge: London).
- Putnam, H. (1975/1985) The meaning of 'meaning'. *Philosophical Papers, Vol. 2: Mind, Language and Reality* (Cambridge University Press: Cambridge).
- Schank, R. (1972) Conceptual Dependency: A Theory of Natural Language Understanding. *Cognitive Psychology* 3:552-631.
- Sparck Jones, K. (2004) What's new about the Semantic Web? *ACM SIGIR Forum*, 38(2).
- Surowiecki, J. (2004) *The Wisdom of Crowds* (Random House: New York).
- Wilks, Y. (1968) *Computable Semantic Derivations*. Systems Development Corporation, SP-3017.
- Wilks, Y. (1978) Programs and Texts. *Computers and the Humanities* 11:134-149.
- Wilks, Y. (2004) Companions: a new paradigm for agents. *Proc. International AMI Workshop, IDIAP, Martigny, CH.*
- Wilks, Y. (2005) What would a Wittgensteinian Computational Linguistics be like? *Proc. International Congress on Pragmatics, Garda, Italy.*
- Wittgenstein, L. (1953) *Philosophical Investigations* (Oxford University Press: Oxford).
- Woods, W. (1975) What's in a Link: Foundations for Semantic Networks. *Representation and Understanding: Studies in Cognitive Science* (Academic Press: New York), pp. 35-82.