

The Semantic Web: Apotheosis of Annotation, but What Are Its Semantics?

Yorick Wilks, *University of Sheffield*

This article discusses what kind of entity the proposed Semantic Web is, principally by reference to the relationship of natural language structure to knowledge representation.

In the middle of a cloudy thing is another cloudy thing, and within that another cloudy thing, inside which is yet another cloudy thing ... and in that is yet another cloudy thing, inside which is something perfectly clear and definite."—ancient Sufi saying

This article considers what kind of object the Semantic Web (SW) is to be. In particular, it asks about SW semantics in the context of the relationship between knowledge representations (KRs) and natural language. This is a vast, and possibly ill-formed, issue, but the SW is no longer simply an aspiration in a magazine article;¹ it's a serious

research subject worldwide, with its own conferences and journal. So, although the SW might not yet exist in a demonstrable form, in the way the Web itself plainly does, it's a topic about which we can ask fundamental questions as to its representations, their meanings, and their groundings, if any.

The concept of the SW has two distinct origins, and this bifurcation persists in two differing lines of SW research: one closely allied to notions of documents and natural language processing (NLP) and one not. These differences of emphasis or content carry with them different commitments about what it is to interpret a KR and what the interpretation method has to do with meaning in natural language.

I'll try to explore both these strands here, but my assumptions will be consistent with the first branch of the bifurcation. That is, I assume that natu-

ral language is, in some clear sense, humans' primary method of conveying meaning and that other methods of conveying meaning (formalisms, science, mathematics, codes, and so on) are parasitic upon it. This view isn't novel: it was once associated firmly with the philosophy of Ludwig Wittgenstein,² who I believe is slightly more relevant to these issues than Graeme Hirst argued with his immortal, satirical, line,

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.³

The quotation at the beginning of this article is intended to suggest not a skeptical position but one where the SW will become a reality. Many popular criticisms of the SW (for example, see [MAY/JUNE 2008](http://</p>
</div>
<div data-bbox=)

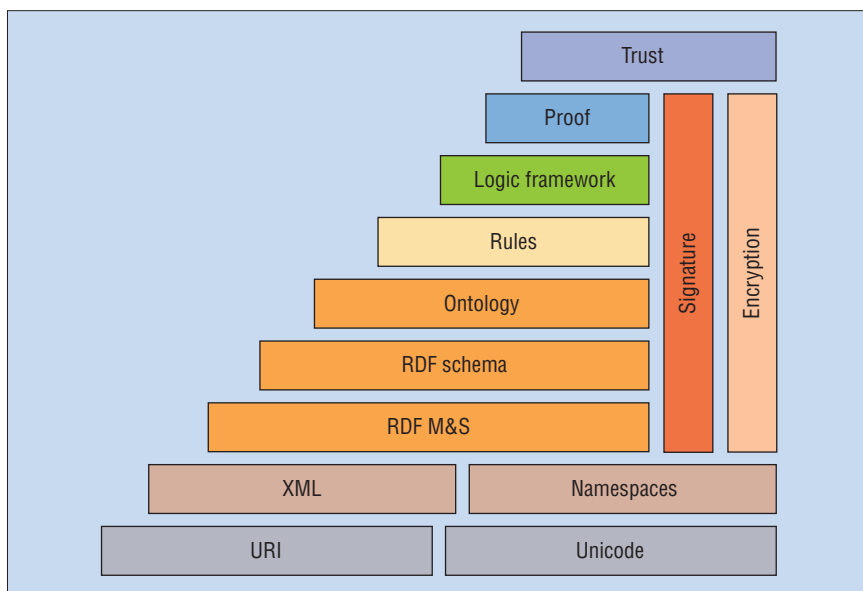


Figure 1. Levels of annotation and objects in the Semantic Web.¹ If you look at only the upper levels, the SW appears to be just another instance of “Good Old-Fashioned AI.” However, the lower levels (namespaces, XML, and RDFS) are products in part of natural language processing.

halfanhour.blogspot.com/2007/03/why-semantic-web-will-fail.html) don't examine foundational issues with any care. Moreover, they fail to see that their thrust—for example, that agreed ontologies in a field are difficult to obtain—implies that science and medicine can't be formalized at all, quite independently of the SW's existence. Such a view is completely at odds with current developments in e-science practice,⁴ and indeed the whole history of science itself.

The Semantic Web and AI

Hirst's comment serves to show that any relation between philosophies of meaning, such as Wittgenstein's, and classic AI (or GOFAI—Good Old-Fashioned AI—as it's often known) isn't an easy one. GOFAI remains committed to some form of logical representation for the expression of meanings and inferences, even if it isn't the standard forms of the predicate calculus. Most issues of *AI Journal* consist of papers in this genre.

Some have taken the initial presentation of the SW by Tim Berners-Lee, James Hendler, and Ora Lassila¹ to be a restatement of the GOFAI agenda in new and fashionable WWW terms. Their article describes a system of services, such as scheduling a doctor appointment for an elderly relative, that would require planning, accessing the databases of both the doctor's and relative's diaries, and synchronizing those databases.

Such planning behavior has been at the heart of GOFAI, and there has been a direct transition (quite outside the discussion of the SW by some researchers) from decades of research on formal KR in AI to the modern discussion of ontologies. This is clearest in work on formal ontologies representing the content of science,^{5,6} where many of the same researchers have transferred discussion and research from one paradigm to the other.

All this has been done under what you could call the standard KR assumption in AI. This assumption goes back to the earliest research on systematic KR by John McCarthy and Patrick Hayes,⁷ which we could consider as defining core GOFAI. The assumption here is that the predicates in such representations merely look like English words but are in fact formal objects, loosely related to the corresponding English but without its ambiguity, vagueness, and ability to acquire new senses with use. This assumption has been apparent in both the original SW paper and some of what has flowed from it, and I shall return to it later.

But few of the complex theories of KR in GOFAI (McCarthy and Hayes' fluents,⁷ McCarthy's later autoepistemic logic,⁸ Hayes' “naïve physics,”⁹ and Daniel Bobrow and Terry Winograd's Knowledge Representation Language,¹⁰ to name but a few prominent examples) have appeared in SW con-

tributions. A continuity of goals between GOFAI and the SW hasn't meant continuity of research traditions; this is both a gain and a loss. We've gained simpler representation schemes that are probably computable. The loss is due to the lack of sophistication in current schemes of the DAML+OIL (www.w3.org/TR/daml+oil-reference) family and whether they now have the representational power to handle the complexity of the commonsense or scientific world, a point I return to later.

There have been at least two other traditions of input to what we now call the SW, and I'll discuss one in some detail: the way in which the SW concept has grown from the traditions of document annotation.

Annotation and the SW's lower end

Looking at the classic SW diagram from the original *Scientific American* article (see Figure 1), the tendency is to focus on the upper levels: rules, logic framework, and proof. It's these, and their traditional interpretations, that have caused both the SW's critics and admirers to say that it's GOFAI by another name. But looking at the lower levels, you find namespaces and XML, which are the products of what we can broadly call NLP. These products stem from the annotation of texts by a range of NLP technologies we can conveniently gather under the name *information extraction* (IE).¹¹

The available information for science, business, and everyday life still exists overwhelmingly as text; for example, 85 percent of business data is unstructured data (that is, text). This is also true of the Web, although the proportion of it that's text is almost certainly decreasing. And how can the Web be absorbed into the SW except by extracting information from natural text and storing it in some other form—for example, facts stored in a database or text annotations stored as metadata either with or separate from the texts themselves? These forms are exactly those that large-scale IE provides.¹² If, on the other hand, we take the view that the Web won't become part of the SW, we face an implausible evolutionary situation of a new structure starting up with no reference to its vast, functioning, but more primitive predecessor. Things just don't happen like that.

XML, the annotation standard that has fragmented into a range of cognates for particular domains (for example, TimeML and VoiceML), is only the latest standard in

the history of annotation languages. These languages attach codings to individual text items to indicate information about them or what should be done with them in some process, such as printing. Indeed, annotation languages originated partly as metadata for publishing documents (the Stanford roff languages, then Donald Knuth's TeX, and later LaTeX), as well as semi-independently in the humanities community as a way to formalize scholarly annotation of text. The Text Encoding Initiative adopted SGML (Standard Generalized Markup Language), a development of Charles Goldfarb's original GML.¹³ SGML in turn became the origin of HTML (as a proper subset), which then gave rise to XML as well as being the genesis of the NLP annotation movement that initially underpinned IE technology.

There were early divisions over exactly how and where to store text annotations for computational purposes. For example, in SGML, annotations were infix in the text with additional characters (as in LaTeX), which made the annotated text more difficult for humans to read. The DARPA research community, on the other hand, produced a functioning IE technology that stored annotations (indexed by spans of characters in the text) separately as metadata. This tradition is preserved in the University of Sheffield's GATE (General Architecture for Text Engineering) language-processing platform,¹² for example, and underpins many European SW projects.^{14,15} This was one of the two origins of the metadata concept, the other being the index terms that were the basis of the standard information-retrieval (IR) approach to document relevance.

IE technology has some 25 years of history, which began with the hand-coded approaches of Naomi Sager¹⁶ and Gerald DeJong.¹⁷ IE then moved to a fully automatic system with tools such as the CLAWS4 program¹⁸ for part-of-speech tagging. This was the first program that systematically added to a text "what it meant" even at the low level of interpretation that such tags represent. IE now reliably locates names in text and their semantic types, and relates them together by means of learned structures called templates into forms of facts and events. Such structures are virtually identical to the RDF triple stores that form the basis of the SW, which aren't quite logic but are similar to IE output. IE began by simply automating annotation but has progressed to

the point where "annotation engines" based on machine learning¹⁵ can learn to annotate in any form and in any domain.

Extensions of IE technology have led to effective question-answering systems trained from text corpora in well-controlled competitions and, more recently, to the use of IE patterns to build ontologies directly from texts.¹⁹ Ontologies are basically conceptual-knowledge structures, which organize facts derived from IE at a higher level. They're close to the traditional KR goal of AI and occupy the middle level in the original SW diagram. I'll return to them later. My point here is just that the SW inevitably rests on some technology within the scope of IE, to annotate raw texts to derive company and

The available information
for science, business,
and everyday life still exists
overwhelmingly as text;
for example, 85 percent
of business data is text.

person names first, then semantic typings of entities, then fact databases, and later ontologies. Where would lists of names, and namable objects, come from, if not automatically from texts? Are we to imagine that researchers make up such inventories?

This view of the SW underlies most European work on the SW and Web services.²⁰ In this view, the SW at its base level is a conversion from the Web of texts through an annotation process of increasing grasp and vision. Such a process projects notions of meaning up the classic SW diagram from the bottom. Richard Braithwaite wrote a classic book on how scientific theories get the semantic interpretation of "high level" abstract entities (such as neutrinos or bosons) from low-level data.²¹ He called this process *semantic ascent* up a hierarchically ordered scientific theory. This view of the SW, which sees NLP and IE among its foundational processes, bears a striking resemblance to that view of scientific theories in general.

Blurring the text-program distinction

These IE technologies add "the meaning of a text" to Web content in varying degrees and forms. They also constitute a blurring of the distinction between language and KR, because the annotations are themselves forms of language, sometimes close indeed to the language they annotate. This process at the same time blurs the distinction between programs and language itself. Historically, two contrary assertions have already blurred this distinction:

- Texts are really programs (which is one form of GOFAT).
- Programs are really texts.

As to the first assertion, there's Carl Hewitt's claim that "language is essentially a side effect" in AI programming and knowledge manipulation.²² H. Christopher Longuet-Higgins argued that English was essentially a high-level programming language.²³ Edsger Dijkstra's view of natural language (personal communication) was essentially that natural languages weren't up to the job they had to do and would be better replaced by precise programs, which is close to being a form of this assertion.

A smaller group—what you might term the "Wittgensteinian opposition"—maintains the second assertion. From this group's perspective, natural language is and always must be the primary KR device. As I mentioned before, all other representations, no matter what their purported precision, are parasitic upon language; they couldn't exist if language didn't.²⁴ The reverse isn't true, of course, and hasn't been for most of human history. Such representations can never be wholly divorced from language, in terms of their interpretation and use. This article is intended as a modest contribution to that tradition; a great deal more can be found in a dialogue with Sergei Nirenburg.²⁵

According to this second perspective, systematic annotations are just the most recent bridge from language to programs and logic. Not long ago, it was perfectly acceptable to assume that a KR must be derivable from an unstructured form—that is, natural language. As William Woods stated,

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.²⁶

The emphasis here is on a method of going from the less to the more formal, a process that inevitably imposes a dependency between the two representational forms (language and logic). This gap has opened and closed in different research periods. In the original McCarthy and Hayes writings on KR in AI,⁷ it's clear, as with Hewitt and Dijkstra's views, that they thought language was vague and dispensable. We can view the annotation movement associated with the SW as closing the gap in the way in that Woods described.

The separation of annotations into metadata has strengthened the view that the original language from which the annotation was derived is dispensable. However, the infixing of annotations in a text suggests that the whole (original plus annotations) still forms some kind of object. The "dispensability of the text" view doesn't depend on the type of representation derived—in particular, logical or quasilogical representations. Roger Schank considered the text dispensable after his Conceptual Dependency representations had been derived. This is because he believed that those representations contained the text's whole meaning, implicit and explicit, even though they wouldn't be considered any kind of formal KR.²⁷ This is a key issue that divides opinion here: Can we know that any representation whatsoever contains all and only the meaning content of a text, and what would it be like to know that?

Standard philosophical problems such as this one might or might not 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 would likely have castigated all such annotations as "markerese," his name for any markup coding with objects still recognizably in natural language and thus not reaching to any meaning outside language.²⁸

The SW movement, at least as I've described it here, takes this criticism head on and continues onward, hoping that URIs and what some call the "popping out of the virtual world" (for example, by giving a Web representation your concrete phone number) will solve semantic problems. That is, this movement accepts that the SW, even if it's based on language via annotations, will provide sufficient inferential traction with which to run Web services.

Is this plausible? Can all you want to

know be put in RDF triples, and can they then support the subsequent reasoning required? Even when agents thus based seem to work in practice, nothing will satisfy a critic such as Lewis except a Web based on a firm (that is, formal and extrasymbolic) semantics and effectively unrelated to language at all. But a century of experience with computational logic has shown that this can't be had outside narrow and complete domains. So, the SW might be the best way of showing that a nonformal semantics can work effectively, just as language itself does, and in some of the same ways.

An IR critique of SW semantics

In a critique of the SW, Karen Spärck Jones

It might now be possible,
using the whole Web—
and thus reducing data
sparsity—to produce
much larger models
of a language.

returned to a theme she had deployed before against much non-empirically based NLP such as ontology building: "words stand for themselves" and not for anything else.²⁹ This claim has been the basis of successful IR research in the Web and elsewhere. Content, for her, can't be recoded in any general way, especially if it's general content as opposed to that from a specific domain. In a specific domain, such as medicine, she seemed to believe technical ontologies might be possible as representations of content. As she put it mischievously, 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, and primitives). This task is what automated annotation attempts to do on an industrial scale. Spärck Jones' key example is (in part) "A Charles II parcel-gilt cagework cup, circa 1670." What, she asks, can be recoded there, into

any other formalism, beyond the relatively trivial form {object type: CUP}?

What, she asks, of the rest of that (perfectly real and useful) description of an artifact in an auction catalog, can be rendered other than in the exact words of the catalog (and of course their associated positional information in the phrase)? This is a powerful argument, even though this example's persuasiveness might rest more than she would admit on it being one of a special class of cases. The fact remains that content can in general be expressed in other words: this is what dictionaries, translations, and summaries routinely do. Where she's right is that GOFAI researchers are wrong to ignore the continuity of their predicates and classifiers with the language words they clearly resemble, and often differ from only by being written in uppercase.²⁵ What can be done to ameliorate this impasse?

One method is to construct empirical ontologies from corpora,^{19,30} now a well-established technology, even if it can't yet create complete ontologies. This is a version of the previous Woods quote, according to which a KR (an ontological one in this case) must be linked to some natural language text to be justifiably derived. We can then consider this derivation to give meaning to the conceptual classifier terms in the ontology, in a way that just writing them down a priori doesn't.

An analogy here would be with grammars. When linguists wrote them down "out of their heads," those grammars 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 produce better structures from sentences by using those rules.

A second method for dealing with the impasse is to return to the observation that we must take "words as they stand."²⁹ But perhaps, to adapt George Orwell, not all words are equal; perhaps some are aristocrats, not democrats. From this perspective, what were traditionally called semantic primitives remain just words but are also special words: a set that form a special language for translation or coding, albeit one whose members remain ambiguous, like all language words. If such privileged words exist, perhaps we can have explanations and innateness (even definitions) alongside an empiricism of use. John Olney showed that counts of the words used in definitions in an actual dictionary (*Webster's Third New In-*

ternational Dictionary, in his case) reveal a clear set of primitives on which all the dictionary's definitions rest.³¹

By "empiricism of use," I mean the approach that has been standard in NLP since the work of Frederick Jelinek and John Lafferty³² and that has effectively driven GOFAI-style approaches based on logic to the periphery of NLP. Jelinek attempted to build a machine translation system at IBM based entirely on machine learning from bilingual corpora. He wasn't ultimately successful, in the sense that his results never beat those from SYSTRAN, the leading handcrafted system. However, he changed NLP's direction; researchers consequently tried to reconstruct, by empirical methods, the linguistic objects on which NLP had traditionally rested: lexicons, grammars, and so on.

The barrier to further advances in NLP by these methods seems to be the data-sparsity problem to which Jelinek originally drew attention. In short, the problem is that language is a system of rare events. A complete model for a language (say, at the trigram level, a trigram being a sequence of three words from a larger sequence of words) seems very difficult to derive. Much of any new, unseen, text corpus may always remain uncovered by such a model.

The Web as a corpus

However, it might now be possible, using the whole Web—and thus reducing data sparsity—to produce much larger models of a language. This could bring us far closer to the full language model necessary for tasks such as complete annotation and automatically generated ontologies. The Wittgensteinian will always want to look for the use rather than the meaning, and nowhere has more use become available than on the whole Web itself. Here, I briefly describe research that attempts to make data for a language much less sparse, without loss. These results are as yet only suggestive and incomplete, but they do seem to offer a way forward.

Adam Kilgarriff and Gregory Grefenstette were among the first to point out 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.³³ A rough computation shows that a person would need about 60,000 years to read all the English documents now on the Web. But the issue here isn't building a psychological model of an individual, so this

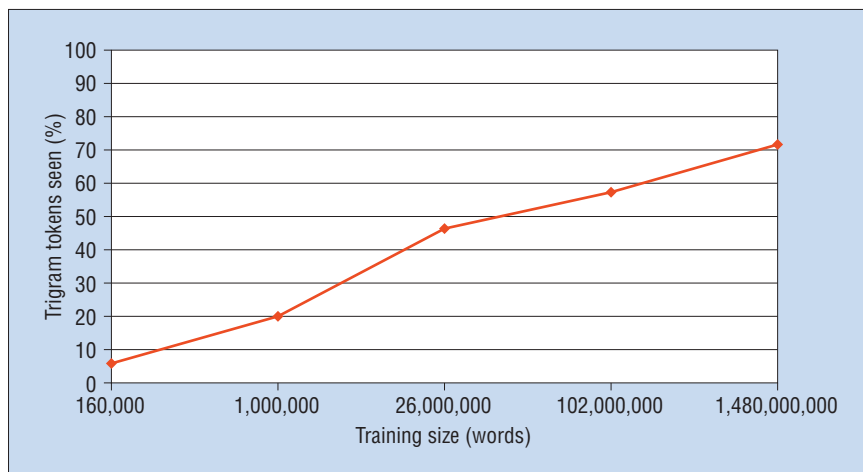


Figure 2. The percentage of trigrams seen with training-corpus size. The percentage grows linearly with the training-corpus size.

fact about size needn't deter us. Roger Moore noted that, if a baby had to use current NLP speech-learning methods, learning to speak would require a hundred years of exposure to data.³⁴ But this fact hasn't deterred the development of effective speech technology. Grefenstette provided a simple and striking demonstration of the value of treating the whole Web as a corpus. His experiments showed that the most frequent translation of a word pair on the Web—from among all possible translation-equivalent word pairs—is invariably the correct one.³⁵

The Reveal project takes large corpora, such as a 1.5-billion-word corpus from the Web, and asks how much of a test corpus is covered by the trigrams in that large training corpus.³⁶ The project considers both regular trigrams and *skip-grams*, which are trigrams consisting of any discontinuity of items with a maximum window of four skips between any of the trigram's members. Consider this sentence:

Chelsea celebrates Premiership success.

The two standard trigrams in that sequence are

Chelsea celebrates Premiership
celebrates Premiership success

But the one-skip trigrams will be

Chelsea celebrates success

Chelsea Premiership success

These skip-grams seem at least as informative, intuitively, as the original trigrams. Reveal experiments suggest that, surprisingly,

skip-grams buy additional coverage without the expense of producing nonsense. Recent research shows that using skip-grams can be more effective than increasing the corpus size. For a 50-million-word corpus, skip-grams have achieved results similar to (in terms of coverage of test texts) those of quadrupling the corpus size. This illustrates the possible use of skip-grams to expand contextual information to get something closer to 100 percent coverage. Such an approach would combine greater coverage with little degradation, thus achieving something much closer to Jelinek's original goal for an empirical corpus linguistics.

The 1.5-billion-word training corpus contained 67 percent of those trigrams appearing in randomly chosen 1,000-word test texts in English. That is, 67 percent of the trigrams found in any random 1,000-word passage of English were found in the gigaword corpus. But my colleagues and I obtained 74 percent coverage with four-skip trigrams (see Figure 2). This suggests, by extrapolation, that to achieve 100 percent trigram coverage (including skip-grams of up to four skips), a corpus must contain 75×10^{10} words. Our corpus giving 74 percent coverage was 15×10^8 words, and Grefenstette calculated there were more than 10^{11} English words on the Web in 2003³⁵ (that is, about 12 times what Google indexed at that time). So, the corpus needed for complete coverage of training texts by trigrams would be about seven times the full English Web in 2003, which is somewhat closer to the size of today's (2007) English Web.

All this is, again, preliminary and tentative, but it suggests that an empiricism of

usage might now be more accessible (with corpora closer to the whole Web) than Jelinek thought at the time (1990) of his major machine translation work at IBM.

Because such modern Web corpora are so vast they can't conceivably offer a model of how humans process semantics, a cognitive semantics based on such usage remains an open question. However, one way forward might be to adapt skip-grams so that they can pick up agent-action-object triples capturing protofacts in very large numbers (perhaps with the aid of a large-scale fast surface parser of the kind already applied to large chunks of the Web). This is an old dream going back at least to 1968, where I viewed similar triples as trivial Wittgensteinian "forms of fact."³⁷ These extracted (triple) text objects were later revived by Kilgarriff and Grefenstette as a "massive lexicon"³³ and are now available as inventories of surface facts at ISL.³⁸ These objects don't differ much from standard RDF triples and might offer a way to cheaply derive massive SW content, even more simply than by machine learning-based IE.

If anything along these lines is possible, then NLP will be able to provide the base semantics of the SW more effectively than it does now, by using a large portion of the Web as its corpus. If you find this notion unattractive, I challenge you to demonstrate some other plausible technique for deriving the massive RDF content the SW will require. Can anyone seriously believe this can be done other than by NLP techniques of the type I've been describing?

A third view of the SW: Trusted databases

This third view emphasizes databases as the SW's core. In this view, a cadre of guardians protects the databases' integrity, keeping the meanings of their features constant and trustworthy. This is a matter quite separate from both logical representations (dear to GOFAT) and any language-based methodology such as I've described in this article.

This view is, I believe, close to Berners-Lee's own vision of the SW.¹ His vision deserves extended discussion and consideration that can't be given here, but it will inevitably suffer from the difficulty of any view (such as GOFAT) that seeks to preserve predicates, features, facets, or whatever from the NLP vagaries of changing sense and drift over time. We still "dial" numbers when we make a phone call, even though

telephones no longer have dials; so not even number-associated concepts are safe from time. The long-running Cyc project,³⁹ one of the predecessors of the SW as a universal repository of formalized knowledge, suffered from precisely such "predicate drift": predicates don't mean this year what coders meant by them 20 years earlier. The SW at present offers no solution to this problem.

Berners-Lee's vision has the virtues and defects of Hilary Putnam's later theory of meaning, where scientists become the guardians of meaning.⁴⁰ For example, only scientists know the true chemical nature of molybdenum and how it differs from aluminum, which has the same appearance. So, only these guardians know the *meaning*

We still "dial" numbers
when we make a phone call,
even though telephones
no longer have dials; so not
even number-associated
concepts are safe from time.

of molybdenum. Putnam's theory required that scientists don't allow the criteria of meaning to leak to the general public, lest the criteria become subject to change. Many observers have argued that you can't make this separation, in principle or in practice, because scientists are only language users in lab coats.^{41,42}

The representation of tractable scientific knowledge

For a concrete illustration of issues raised by the scientific-database view of the SW, we can consider the questions of meaning and interpretation of formal knowledge that Toni Kazic first asked in connection with biological databases. These questions could be expected to form part of any SW wide enough to cover scientific and technical knowledge. Kazic has raised a number of issues close in spirit to those of this article,⁴³ but against a background of expert knowledge of biology that would be hard to capture here.

In brief, she draws attention to two symmetric chemical reactions of cleavage (a molecule splitting into simpler molecules), which we can write as $A \leftrightarrow B$ and $C \leftrightarrow D$. An enzyme Z catalyzes both these reactions, according to KEGG (Kyoto Encyclopedia of Genes and Genomes, www.genome.jp/kegg/kegg1.html) maps, the standard knowledge structures in the field. However, catalyzing these two reactions is normally the province of Y compounds. Z isn't in class Y, so it shouldn't, in standard theory, be able to catalyze the reactions. Yet it does. A comment in the KEGG maps states that Z can catalyze reactions such as those of another enzyme Z' under some circumstances, where Z' actually is a Y, although its reactions differ considerably from Z. In addition, Z and Z' can't 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?"⁴³

This is an inadequate attempt to state the biological facts in this nonspecialist form, but it's clear that something odd is going on here, something that Marxists might once have hailed as a dialectical or contradictory relationship. It's certainly an abstract structure that challenges conventional KR's. It's also far more complex than the standard form of default reasoning in AI, which takes the view that if anything is an elephant, it has four legs, even though Clyde, undoubtedly an elephant, has only three.

The flavor of the phenomena here is that of extreme context dependence—that is to say, an entity behaves quite differently—indeed in opposite fashions—in the presence of certain other entities. Languages are, of course, full of such phenomena, such as when "cleave to the Lord" and "cleave a log" mean exactly opposite things. We have structures in language representation for describing and representing such phenomena, although there's no reason at the moment to believe they're of any assistance here.

Kazic is making the point that any SW that represents biological information (and licenses correct inferences) must be able to deal with phenomena as complex as this. At first sight, such phenomena seem beyond the ability of a standard ontology dependent on context-free relations of inclusion and

the other standard relations, as Kazic puts this matter:

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.⁴³

In connection with the initial translation into RDF, she continues,

Building a tree of phrases to emulate binding ... forces one to say explicitly something one may not know (for example, 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.⁴³

This quotation is revealing about the structure of science and the degree to which it remains partly a craft skill, even in the most technical modern areas. Even if that weren't the case, being forced to be more explicit and to remove ambiguities could have only a positive influence. The quotation brings out the dilemma in some parts of advanced science that intend to use the SW: whether the science is yet explicit enough and well understood enough to be formally coded. This question is quite separate from issues of whether the proposed codings (from RDF to DAML+OIL) have the representational power to express what's to be made explicit.

If biology isn't yet explicit and well-enough understood, then it might not be so different from ordinary life as we might have thought. It's certainly not so different from the language of auction house catalogs, as in Spärck Jones' example. In that example, the semantics remains implicit, in that it rests on our human interpretation of the words of annotations or comments (or, in Kazic's case, in the margins of KEGG maps).

The analogy here isn't precise, of course: current SW representational styles have, to some degree, sacrificed representational sophistication to computational tractability (as, in a different way, the Web itself did in the early '90s). Perhaps, when some of the greater representational powers in traditional GOFAI research are brought to bear, the KEGG-style comments might be translated from English phrases with an implicit semantics to the explicit semantics of ontologies and rules. This is what we must all hope for. But in the case of Spärck Jones' description of the 17th-century cup, the problem doesn't lie in any KR. It lies only

in the fact that the terms involved are all so precise and specific that no generalizations—no imaginable auction ontology—would provide a coding that lets us throw away the original English. The possibility always remains of translation into another language or an explicit numbering of all the concepts in the passage, but neither route provides any representational savings.⁴⁴

Kazic goes on to argue that one effect of these difficulties about explicitness is that “most of the semantics are pushed onto the applications,”⁴³ where the Web agents might or might not work, but there's insufficient explicitness to know why in either case. This is a traditional problem. For example, a major AI objection to the connectionist/

If biology isn't yet explicit
and well-enough understood,
then it might not be
so different from ordinary
life as we might
have thought.

neural net movement was that, whether the approach worked or not, nothing was served scientifically if what it did wasn't understood—that is, transparent and explicit. We don't yet have enough SW data to be sure, but it's completely against the spirit of the SW that its operations should be unnecessarily opaque or covert. This becomes even clearer if you see the SW as the Web plus the meanings, where you would expect only additional, not less explicit, information.

Discussions in this area normally avoid more traditional ontological inquiries—namely, what things there are in the world. Ancient questions have a habit of returning to bite you at the end, though. In this article, I've taken a robust position, in the spirit of Willard Quine,⁴⁵ that whatever we put into our representations—concepts, sets, and so on—has existence, at least as a polite convention. But a fully explicit SW might have to make ontological commitments of a more traditional sort, at least regarding the URIs—the points where the SW meets

the world of unique descriptions of real things. But scientific examples of this interface in the world of genes are by no means straightforward.

Suppose we ask, what are the ontological objects in genetics—say, in the classic *Drosophila* database FlyBase?⁴⁶ 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, and so on with gene IDs. However, these sequences often need modifying because of new discoveries in the literature. For example, as scientists understand better how genes get expressed in various biological processes, they frequently identify new regulatory regions upstream from the gene sequence. So, the gene ID's “referent” 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 quite happily that the gene “rutabaga” does such-and-such, even if they're aware that rutabaga's referent has changed significantly several times over the last decade. The curators and biologists are, for the most part, content with this situation, although some in the *Drosophila* community have argued that the community overall has been cavalier with gene naming.

This situation, assuming my nonexpert description is broadly correct, shows that ontological issues still exist in the original sense of that word: that is, as to what there actually *is* in the world. More precisely, it directly refutes Putnam's optimistic theory that meaning can ultimately be grounded in science because only scientists know the true criteria for selecting the referents of terms.⁴⁰ The *Drosophila* case shows this isn't so. In some cases geneticists have only a hunch, sometimes proved false in practice, that there are lower-level objects unambiguously corresponding to a gene ID, in the way an elementary molecular structure certainly corresponds to an element's name in Mendeleev's table (and in the way SW URIs correspond to unique data objects).

There's also a fourth view of the SW, one much harder to define and discuss: If the SW just keeps moving as an engineering development and is lucky (as the

successful scale-up of the Web seems to have been luckier, or better designed, than many cynics expected), then real problems won't arise. This view is a hunch and not open to close analysis, but I can only wish it well, without being able to discuss it in detail here. The situation remains that the SW hasn't yet taken off as the Web, Google IR, and iPods did. Maybe something about the SW's semantics is holding it back—something perhaps connected, as I've argued, to its failure so far to generate semiformalized material on a great scale from existing Web material, though this could change at any moment.

NLP will continue to underlie the SW, including its initial construction from unstructured sources such as the Web, whether its advocates realize this or not. Such NLP activity is the only way up to a defensible notion of meaning at conceptual levels (in the original SW diagram) based on lower-level empirical computations of usage. I'm definitely not trying to claim logic-bad, NLP-good in any simple-minded way, but that the SW will be a fascinating interaction of these two methodologies, again like the Web (which has been basically a field for statistical NLP research) but with deeper content.

Only NLP technologies (and chiefly IE) will be able to provide the requisite RDF knowledge stores for the SW from existing Web (unstructured) text databases, and in the vast quantities needed. There is no alternative at this point. A wholly or mostly handcrafted SW is also unthinkable, as is a SW built from scratch and without reference to the Web. I also assume that, whatever the limitations on current SW representational power, the SW will continue to grow in a distributed manner so as to serve scientists' needs, even if it isn't perfect. The Web has already shown how an imperfect artifact can become indispensable.

Contemporary statistical large-scale NLP offers new ways of looking at usage in detail and in quantity, even if we can't easily relate the huge quantities required to an underlying theory of human learning and understanding. We can see glimmerings, in machine learning studies, of something like Wittgenstein's "language games"² in action and of the role of key concepts in the representation of a whole language. Part of this can be done only by some automated recapitulation of primitive concepts' role in the organization of (human-built) ontologies, thesauri, and wordnets.

The heart of the issue is the creation of meaning by some interaction of (unstructured language) usage and the interpretations to be given to higher-level concepts. This is a general issue, but the construction of the SW faces it crucially. This issue could be the critical arena for progress on a problem that goes back at least to Immanuel Kant's classic formulation in terms of "concepts without percepts are empty, percepts without concepts are blind." If we see that opposition as one of language data (such as percepts) to concepts, the risk is of formally defined concepts always remaining empty (see Ian Horrocks and Peter Patel-Schneider's discussions of SW meaning⁴⁷). The answer

The Semantic Web will continue to grow in a distributed manner so as to serve scientists' needs, even if it isn't perfect.

is, of course, to find a way upward from one to the other. ■

Acknowledgments

I'm indebted to many discussions with colleagues in the AKT (Aktive Knowledge Technologies) project at the University of Sheffield and elsewhere, as well as with Arthur Thomas, Christopher Brewster, Ted Nelson, and other colleagues at the Oxford Internet Institute. The passage on *Drosophila* owes a great deal to conversations with Ted Briscoe, but, as always, the errors are my own. This work was funded partly by the Companions project (www.companions-project.org), sponsored by the European Commission as part of the Information Society Technologies program under EC grant IST-FP6-034434.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, May 2001, pp. 34–43.
2. L. Wittgenstein, *Philosophical Investigations*, Oxford Univ. Press, 1953.

3. G. Hirst, "Context as a Spurious Concept," *Proc. Conf. Intelligent Text Processing and Computational Linguistics*, 2000, pp. 273–287.
4. Y. Wilks and M. den Besten, "Digital Technologies Shaping E-research," to be published in *World Wide Science: Promises, Threats, and Realities*, Oxford Univ. Press.
5. P. Patel-Schneider, P.J. Hayes, and I. Horrocks, *OWL Web Ontology: Language Semantics and Abstract Syntax*, W3C recommendation, Feb. 2004, www.w3.org/TR/owl-semantics.
6. I. Horrocks, "Description Logics in Ontology Applications," *Proc. KI/Tableaux 2005*, 2005, www.gusconstan.com/AI/DescriptionLogics.htm.
7. J. McCarthy and P. Hayes, "Some Philosophical Problems from the Point of View of Artificial Intelligence," *Machine Intelligence 4*, D. Michie, ed., Edinburgh Univ. Press, 1969, pp. 463–502.
8. J. McCarthy, *Formalizing Common Sense: Papers by John McCarthy*, Ablex, 1990.
9. P.J. Hayes, "The Naive Physics Manifesto," *Expert Systems in the Micro-Electronic Age*, D. Michie, ed., Edinburgh Univ. Press, 1979, pp. 242–270.
10. D. Bobrow and T. Winograd, "An Overview of KRL, a Knowledge Representation Language," *Cognitive Science*, vol. 1, no. 1, 1977, pp. 3–46.
11. J. Cowie and Y. Wilks, "Information Extraction," *Handbook of Natural Language Processing*, R. Dale, H. Moisl, and H. Somers, eds., Marcel Dekker, 2000, pp. 190–211.
12. H. Cunningham et al., GATE—a TIPSTER-Based General Architecture for Text Engineering," *Proc. TIPSTER Text Program Phase III*, Morgan Kaufmann, 1997, pp. 45–65.
13. C.F. Goldfarb, "SGML: The Reason Why and the First Published Hint," *J. Am. Soc. Information Science*, vol. 48, no. 2, 1997, pp. 44–49.
14. K. Bontcheva and H. Cunningham, "Information Extraction as a Semantic Web Technology: Requirements and Promises," *Proc. Adaptive Text Extraction and Mining Workshop*, 2003, pp. 222–235.
15. F. Ciravegna, "Designing Adaptive Information Extraction for the Semantic Web in Amilcare," *Annotation for the Semantic Web*, S. Handschuh and S. Staab, eds., IOS Press, 2003, pp. 23–45.
16. N. Sager, "The String Parser for Scientific Literature," *Natural Language Processing*, R. Rustin, ed., Cambridge Univ. Press, 1973, pp. 61–87.
17. G. DeJong, "Skimming Stories in Real Time: An Experiment in Integrated Understanding," PhD thesis, Computer Science Dept., Yale Univ., 1979.
18. G. Leech, R. Garside, and M. Bryant, "CLAWS4: The Tagging of the British National Corpus," *Proc. 15th Int'l Conf.*

The Author

Yorick Wilks is a professor of AI at the University of Sheffield and runs the natural-language-processing research group there. He's also a senior research fellow at the Oxford Internet Institute. His research interests are computational pragmatics, computational lexicons, and information extraction. Wilks received his PhD in philosophy and computing from Cambridge University. He's a fellow of the AAAI and the European Coordinating Committee for Artificial Intelligence, a member of the UK Computing Research Council, and a permanent member of the International Committee on Computational Linguistics. He's also the coordinator of the European Commission's Companions project (www.companions-project.org). Contact him at yorick@dcs.shef.ac.uk.

- Computational Linguistics* (COLING 94), 1994, pp. 622–628.
19. C. Brewster et al., "The Ontology: Chimera or Pegasus," *Proc. Dagstuhl Seminar Machine Learning for the Semantic Web*, 2005, pp. 89–101.
 20. B. Norton, S. Chapman, and F. Ciravegna, *Orchestration of Semantic Web Services for Large-Scale Document Annotation*, Springer, 2005.
 21. R. Braithwaite, *Scientific Explanation*, Cambridge Univ. Press, 1956.
 22. C. Hewitt, "Procedural Semantics," *Natural Language Processing*, R. Rustin, ed., Algorithmics Press, 1972, pp. 99–118.
 23. H. Longuet-Higgins, "The Algorithmic Description of Natural Language," *Proc. Royal Soc. London B*, vol. 182, 1972, pp. 255–276.
 24. Y. Wilks, "What Would a Wittgensteinian Computational Linguistics Be Like?" *Proc. 10th Int'l Congress Pragmatics*, 2005, pp. 212–227.
 25. S. Nirenburg and Y. Wilks, "What's in a Symbol," *J. Theoretical and Empirical AI*, vol. 13, no. 1, 2001, pp. 9–23.
 26. W. Woods, "What's in a Link: Foundations for Semantic Networks," *Representation and Understanding: Studies in Cognitive Science*, Academic Press, 1975, pp. 35–82.
 27. R. Schank, "Conceptual Dependency: A Theory of Natural Language Understanding," *Cognitive Psychology*, vol. 3, no. 4, 1972, pp. 67–88.
 28. D. Lewis, "General Semantics," *The Semantics of Natural Language*, D. Davidson and G. Harman, eds., Kluwer, 1972.
 29. K. Spärck Jones, "What's New about the Semantic Web? Some Questions," *ACM SIGIR Forum*, vol. 38, no. 2, 2004, www.sigir.org/forum/2004D/sparck_jones_sigirforum_2004d.pdf.
 30. C. Brewster, F. Ciravegna, and Y. Wilks, "Knowledge Acquisition for Knowledge Management," *Proc. ICAI 2001 Workshop Ontology Learning*, 2001, pp. 121–134.
 31. J. Olney, C. Revard, and P. Ziff, "Some Monsters in Noah's Ark," research memo, Systems Development Corp., 1968.
 32. F. Jelinek and J. Lafferty, "Computation of the Probability of Initial Substring Generation by Stochastic Context-Free Grammars," *Computational Linguistics*, vol. 17, no. 3, 1991, pp. 315–323.
 33. A. Kilgarriff and G. Grefenstette, eds., special issue on the Web as corpus, *Computational Linguistics*, vol. 29, no. 3, 2003.
 34. R.K. Moore, "A Comparison of Data Requirements for ASR Systems and Human Listeners," *Proc. EUROSPEECH 2003*, 2003, pp. 238–251.
 35. G. Grefenstette, "The Scale of the Multilingual Web," presentation at Search Engine Meeting 2004.
 36. D. Guthrie et al., "A Closer Look at Skipgram Modelling," *Proc. 5th Int'l Conf. Language Resources and Evaluation* (LREC 06), European Language Development Assoc., 2006, pp. 101–111.
 37. Y. Wilks, *Computable Semantic Derivations*, tech. report SP-3017, Systems Development Corp., 1968.
 38. E. Hovy, "Key toward Large-Scale Shallow Semantics for Higher-Quality NLP," *Proc. 12th PAACLING Conf.*, 2005, pp. 98–107.
 39. D. Lenat, "CyC: A Large-Scale Investment in Knowledge Infrastructure," *Comm. ACM*, vol. 38, no. 11, 1996, pp. 33–38.
 40. H. Putnam, "The Meaning of 'Meaning,'" *Philosophical Papers, Vol. 2: Mind, Language and Reality*, Cambridge Univ. Press, 1975/1985, pp. 215–271.
 41. D.H. Mellor, "Natural Kinds," *British J. Philosophy of Science*, vol. 28, no. 2, 1977, pp. 1–23.
 42. Y. Wilks, "Putnam and Clarke and Mind and Body," *British J. Philosophy of Science*, vol. 26, no. 3, 1975, pp. 23–30.
 43. T. Kazic, "Putting the Semantics into the Semantic Web: How Well Can It Capture Biology?" *Proc. Pacific Symp. Biocomputing*, 2006, pp. 140–151, <http://helix-web.stanford.edu/psb06/kazic.pdf>.
 44. D. McDermott, "Artificial Intelligence Meets Natural Stupidity," *Mind Design*, J. Haugeland, ed., Bradford, 1981, pp. 143–160.
 45. W.V.O. Quine, *From a Logical Point of View*, Harvard Univ. Press, 1953.
 46. A. Morgan et al., "Gene Name Extraction Using FlyBase Resources," *Proc. ACL Workshop Language Processing in Biomedicine*, Assoc. Computational Linguistics, 2003, pp. 23–41.
 47. I. Horrocks and P. Patel-Schneider, "Three Theses of Representation in the Semantic Web," *Proc. 12th Int'l World Wide Web Conf. (WWW 03)*, ACM Press, 2003, pp. 39–47, www2003.org/cdrom/papers/refereed/p050/p50-horrocks.html.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.

ADVERTISER/PRODUCT INDEX MAY/JUNE 2008

Advertiser

MIT Press

PAGE

4

Advertising Personnel

Sandy Brown, Business Development Manager
phone +1 714 821 8380
fax +1 714 821 4010
sbrown@computer.org

Onkar Sandal, Sales Representative
phone +1 785 843 1234 x218
fax +1 785 843 1853
osandal@allenpress.com

For production information and conference and classified advertising, contact

**Marian Anderson
IEEE Intelligent Systems**
phone +1 714 816 2139
fax +1 714 821 4010
manderson@computer.org