The Semantic Web

Wendy Hall & Kieron O'Hara

Version of article for Robert Meyers (ed.), *Encyclopedia of Complexity and System Science*, Springer

Intelligence, Agents, Multimedia Group
School of Electronics and Computer Science
University of Southampton
Highfield
Southampton SO17 1BJ
United Kingdom
{wh,kmo}@ecs.soton.ac.uk

Article outline

Glossary

Dereferencing

Dereferencing a URI is using the URI to identify the object or resource it refers to.

Folksonomy

'Folksonomy' is a neologism applied to structures that emerge from the practice of 'tagging' Web content. In some Web 2.0 applications, users can apply a tag (a descriptive term) to content such as a photograph or video clip. The tags need only be meaningful to the individual tagger, but if a large enough number of users tag content, descriptive structures analogous to more formal ontologies can emerge that are meaningful to wide communities.

GRDDL

Gleaning Resource Descriptions from Dialects of Languages (GRIDDL, pronounced 'griddle') is a mechanism for helping bootstrap the Semantic Web, by extracting RDF

from XML documents, using transformations expressed in XSLT. GRDDL became a W3C recommendation in 2007.

Metadata

Metadata is data about data. In the context of the Semantic Web, metadata are also called ‘markup’ or ‘annotations’. Because one aim of the Semantic Web is to support machine processing of information, metadata are helpful in describing the content of data. For example, metadata attached to a series of numerals could explain that it represents a zip code or a height or a population figure.

Ontology

An ontology defines, describes and constrains the concepts and relationships that are used in some particular domain of knowledge. Ontologies may therefore have an important role in data sharing, for example by providing a means of expressing which concepts (in the ontology) are referred to by particular terms (in a set of databases, which may use widely differing vocabulary).

OWL

The Web Ontology Language (OWL) is a language for describing and sharing ontologies on the Web. It is an extension of RDF, and has three variants. OWL Full is maximally expressive and compatible with RDF (any legal RDF document is a legal OWL Full document), but is undecidable. OWL DL is based on a series of restrictions of OWL Full which support efficient reasoning, at the cost of losing the strong connection with RDF (not all RDF documents are legal OWL DL documents). OWL Lite is even more restricted in expressivity, but is easier to grasp. OWL became a W3C recommendation in 2004.

RDF

The Resource Description Framework (RDF) is a standard framework for representing data on the Web, representing it in a three-place relation, of subject-relation-object form, called a *triple*. It uses URIs to refer not only to the two items related, but also to the relationship asserted between them. In this way it extends the linking structure of the Web by allowing the nature of the link asserted to be described. Its syntax is XML-based, to support syntactic interoperability between the two. RDF became a W3C recommendation in 1999.

RDF(S)

RDF Schema (RDF(S) or RDFS) is a language for representing information on the Web. It is an extension of RDF to allow the description of the relationships which can be asserted between resources using RDF. So, for instance, RDF allows the assertion of the relationship ‘author’ between, say, ‘Herman_Melville’ and ‘Moby_Dick’, but RDF(S) is required to assert the properties of the ‘author’ relationship (e.g. that every document has at least one author, or that the ‘author’ relationship is the inverse of the ‘written_by’ relationship). RDF(S) became a W3C recommendation in 2004.

Rules and RIF

Rules govern the transformation of, and inference from, data. In particular, when data are being shared, it may be useful to make basic inferences over the data (for example, to determine whether two names refer to the same object or different objects). Rule-

based knowledge such as this cannot be expressed in individual data stores, and may be hard to express in an ontology. The Rule Interchange Format (RIF) is a language, not complete at the time of writing, to express the most common or basic types of rule.

SPARQL

SPARQL is a special query language designed specifically for querying data stored in RDF ('SPARQL' is a recursive acronym standing for 'SPARQL Protocol And RDF Query Language', and is pronounced 'sparkle'). At the time of writing (December 2007), SPARQL was a candidate recommendation by the W3C.

Triples and Triplestores

A triple is a statement in RDF, consisting of a subject, an object, and a binary predicate that relates them. Each item of the triple is identified by a URI reference (or can be a string literal, such as a date or a number or name). A knowledge repository which contains RDF triples is called a triplestore. Triplestores may need to contain several millions of triples, and so need to be able to support fast querying at these potentially very large scales.

URI

A Uniform Resource Identifier (URI) is a string of characters for identifying an abstract or physical object or resource. There are different URI schemes (i.e. different ways to restrict the syntax of a URI to make it meaningful). A Uniform Resource Locator (URL) is a particular type of URI that identifies its object by means of its access mechanism or 'location' in the network. URIs are important in that they act as a standard way of referring to objects on the Web.

W3C

The World Wide Web Consortium (W3C) is an international non-profit consortium which coordinates the development of Web standards, founded in 1994 under the directorship of Sir Tim Berners-Lee. The W3C groups together all the bodies involved in specifying Semantic Web standards into the Semantic Web Activity. A W3C working group works on a standard for a particular formalism, and when the standard is judged by the working group to be in a final form, fit for purpose and properly interoperable with other W3C standards, it ratifies the formalism by recommending it. Hence a W3C recommendation is an important standard.

Web of Data

The Web of Data is another way of referring to or explaining the vision of the Semantic Web, which emphasises the idea of creating links between data rather than documents (as on the current World Wide Web). Linking data, as with linking documents, enables them to be reused in interesting or unexpected contexts. RDF is the anticipated mechanism for creating the links between data; dereferencing one or more of the URIs in an RDF triple will lead to descriptions of the resources referred to, which in turn are likely to contain further triples, which can again lead to dereferencing and so on.

Web Science

Web Science is the multidisciplinary activity of trying to understand the two-way dynamic relationship between Web technology and wider society, in order to ensure that the technological changes made to the Web (including the Semantic Web) are generally beneficial rather than otherwise.

XML

The EXtensible Markup Language (XML) is a language for allowing users to tag or mark up content using tags of their own devising (for instance based on a particular vocabulary used in a small community of practice). As such, XML can serve as a basic data exchange format between applications. It is an advance on other well-known markup languages (such as HTML, a standard language for marking up content for display on the Web) in that it separates instructions to do with content and document structure from those to do with formatting. It is a basic language for representing and exchanging structured information.

Definition

The Semantic Web is a proposed extension to the World Wide Web (WWW) that aims to provide a common framework for sharing and reusing data across applications. The most common interfaces to the World Wide Web present it as a Web of Documents, linked in various ways including hyperlinks. But from the data point of view, each document is a black box – the data are not given independently of their representation in the document. This reduces its power, and also (as most information needs to be extracted from documents by a human agent) inhibits the use of automatic information processing methods on the Web. The Semantic Web is an effort, steered by the World Wide Web Consortium, to develop a set of protocols, formalisms and standards to transform the Web into a *Web of Data*. Links would be between data, and data could be accessed independently of the applications that created them. This would allow both the sharing of data, and the amalgamation of data from different sources, using heterogeneous formats, in new contexts.

Introduction: history, motivation and the development of standards

The idea of the Semantic Web (SW), of exploiting the possibilities for serendipitous reuse of linked data, dates back at least to Sir Tim Berners-Lee's plenary talk at the first International World Wide Web Conference at CERN in Geneva in 1994 [32]. In that talk, Berners-Lee argued that there is too little machine-readable information on the WWW as was currently constituted. "The meaning of the documents is clear to those with a grasp of (normally) English, and the significance of the links is only evident from the context around the anchor. To a computer, then, the Web is a flat, boring world devoid of meaning. This is a pity, as in fact documents on the Web describe real objects and imaginary concepts, and give particular relationships between them. … Adding semantics to the Web involves two things: allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values. Only when we have this extra level of semantics will we be able to use computer power to help us exploit the information to a greater extent than our own reading." Indeed, the original vision of the WWW was intended to support greater machine understanding of people's work and interactions; the 'flat'

understanding of the world that the WWW produces in machines is a first step towards a richer vision.

The SW was from the beginning conceived as a set of layered standards and formalisms (see p.11 below for details). The development of the Resource Description Framework (RDF) in the 1990s was a key technology [79]. RDF is an important formalism, as it allows expression not only *of* the link between two objects, but also about the *nature* of the link itself. Hence, one can follow a chain of links not only via the objects linked, but also the types of links involved. In the WWW of documents, links connect documents written in the Hypertext Markup Language HTML; in the SW, links in RDF connect not only documents but arbitrary things (objects and relationships) identified by the Uniform Resource Identifiers (URIs [36]) in the RDF triples representing the data.

The ability to move between data linked in such a way opens up the possibility of exposing data to the Web, and then being able to access such data from any application. As a simple example, consider personal information such as one's bank statements, information-based resources such as digital photographs, and an application such as a calendar or diary. Each of these depends on data which is controlled by the applications that use them. But in a genuine Web of Data, we could link these data in a productive way – something as simple as being able to present one's financial information in one's calendar. The metadata in one's photographs often includes information about the time of their creation; an application able to get at the photograph metadata and the data on one's calendar might be able to suggest where the photo was taken, and its possible location. The ability to use all these data in a constructive way is impossible without a Web of linked data to enable applications to move between the data sources.

In this way, the SW changes our model of the value of information. Currently, it is generally presumed that the value of information stems from its *scarcity* – people and organisations gain value from information they have gathered, and are given monopoly rights to exploit that information via such legal contrivances as copyright, intellectual property rights, licensing, and so on. Even when organisations do not resort to the law, they will make great investments in protecting trade secrets. However, this scarcity-based model seems inadequate for the digital age.

In the first place, as economist William Baumol has argued, the social benefits from unlicensed use of 'protected' knowledge and innovation, were already large in the pre-digital economy, and indeed account for much of our wealth today: "some 80 percent of the benefits [from innovation] may plausibly have gone to persons who made no direct contribution to innovation. The rather startling implication of all this is that the spillovers of innovation, both direct and indirect, can be estimated to constitute well over half of current [US] GDP – and it can even be argued that this is a very conservative figure" [31, p.135]. And secondly, the Internet and the Web have made it harder to preserve monopoly rights to information, as copying and distribution reduces the marginal cost to producers to close to zero. Although many media companies have taken rearguard action to protect their intellectual property, so simple is the distribution model on the Web that the basis of the value of information is rapidly switching from scarcity to *abundance*. It is the large quantity of data, that can be placed in novel and unintended contexts with little cost, that makes it increasingly valuable in the age of digital technologies – and it is this abundance that the SW is designed to foster.

One of the major drivers of the Web of Data has been the transformation of science into *e-science*, a computer-enabled, data-heavy view of science as the analysis of the very large quantities of information that improved instrumentation, larger computer power, more prevalent sensor networks and greater memory storage have released. Several disciplines have seized on the opportunity to exploit such data, which are often available only in diverse and heterogeneous datasets. In particular, interdisciplinary research is growing in importance, requiring data developed in different disciplines, using a confusion of vocabularies and methods of collection. Methods for dealing with such large and heterogeneous datasets are required in many areas, including the life sciences, climate research, medicine and epidemiology, and genomics, to name but four, which explains the interest in many of these fields in the SW.

The use of the SW in such large, public projects was perhaps predictable, but much debate and discussion has focused more on the individual's interface to it. We discuss this in more detail below, but one reason for this was that the landmark publication for the public view of the SW, an article in *Scientific American* for 2001, written by Berners-Lee, James Hendler and Ora Lassila [39], developed the idea of a Web of Data with a number of household gadgets interfacing with it. The possibilities envisaged in their scenario included: a telephone that turned down the volume of all local devices with volume controls when it rang; an agent that could plan a programme of medical care; and a calendar that could integrate this information to adjust a set of appointments. The point of the article was not the impressive set of agents, but rather the Web of Data that sat underneath them. However, many readers focused on the gadgets, and – given that such gadgets are not at the time of writing very common or effective – have concluded either that the SW has been a failure, or that it was an unrealistic vision from the beginning (see 'Controversies' below, p.20).

In 2006, a further publication by Berners-Lee, together with Nigel Shadbolt and Wendy Hall, appeared in the publication *IEEE Intelligent Systems* de-emphasising the agents and focusing on the idea of the SW as a Web of Data or actionable information [100]. This paper argued that the agents described in 2001 could only flourish when standards for data sharing are well-established. The need for such standards, and for SW technologies in general, was growing, thanks to developments such as e-science, information-based medicine, and e-government.

Since the late 1990s, the World Wide Web Consortium (W3C), under the direction of Berners-Lee, has led the drive to create the standards. Figure 1 shows a diagrammatic representation of the progress of the SW in the development of its layered standards (the hierarchically-arranged layers are marked in Figure 1 down the left hand column as markup/data/ontology etc). The development of each layer goes through a long, sometimes tortuous, process of research and discussion. Initial stages of research, sometimes competitive, after some time produce a rough consensus about the general properties of a formalism to implement a layer. At that point, consideration is undertaken about creating a Web standard by the W3C.

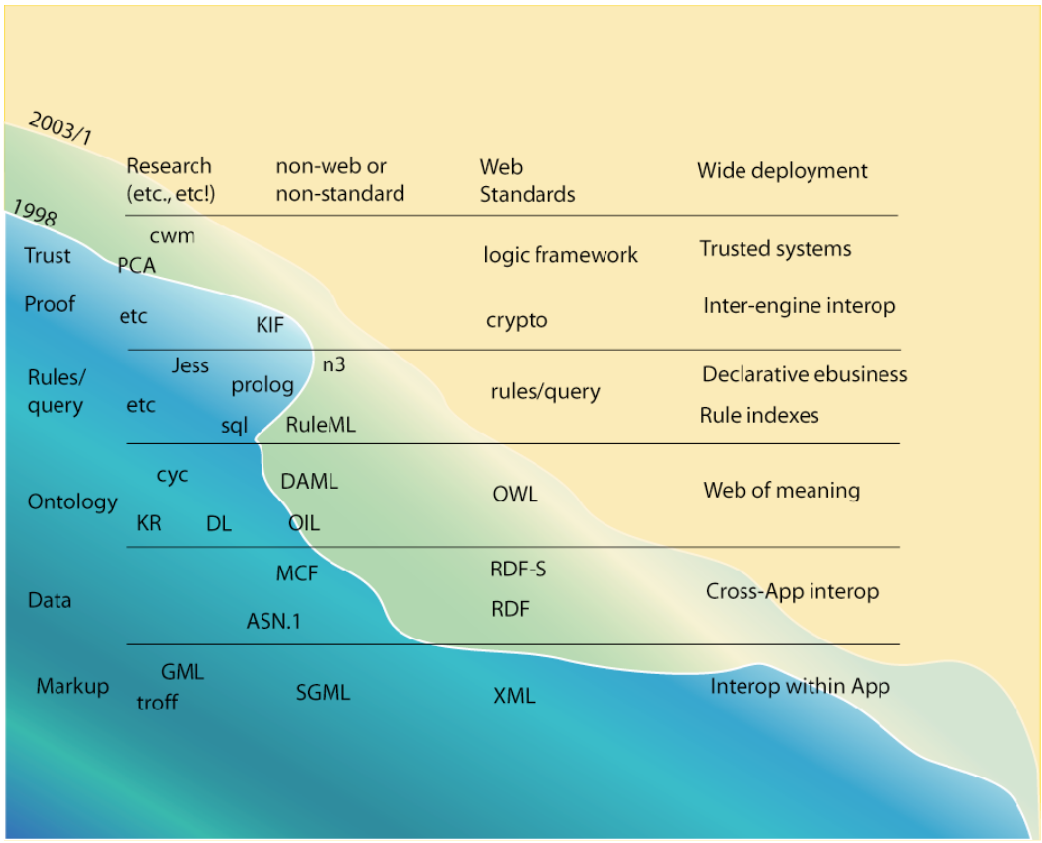

Figure 1: The wave of development of the Semantic Web [50]

To create a standard, a group, representative of stakeholders from academe and private enterprise, is assembled by the W3C, which first creates a *working draft*, which is released under that status for review by the community. Commentary is welcomed, and the draft may be changed dramatically. Once the group responsible for the standard is satisfied that it is fully capable of doing what it required, it is released as a *candidate recommendation*, when it is critiqued in terms of the practicability of its implementation. The next stage is to become a *proposed recommendation*, when it is submitted to the W3C advisory council. Finally, it is released as a W3C recommendation. It generally takes years to negotiate these various stages. Currently, the markup language XML, the data representation and interchange languages RDF and RDF(S), and the ontology language OWL are full recommendations. The next layer up from the ontology layer is that of rules and querying: the proposed query language SPARQL is a proposed recommendation at the time of writing [95], while the rule expression language RIF is at the time of writing at the working draft stage [40].

Figure 1 represents the historical progress of the SW in terms of a ‘wave’ rolling over ‘dry land’. The depth of the ‘sea’ indicates the extent to which SW standards have been accepted and been deployed widely. The SW wave is travelling from the bottom left of the diagram to the top right; the net result of this is that the lower levels of the SW are coming into being and wide acceptance ahead of the higher levels. RDF is, at the time of writing, reasonably widely deployed; SPARQL is nearing the point at which it is becomes an official W3C recommendation; trust (the highest level represented here) is still a research issue, and the consensus about what form a standard to represent and promote trust in data is still in the process of formation.

Linking data

The underlying aim of the SW is to allow data to be explored and queried on the Web, analogously to the way that documents are currently investigated online. One precondition for this is obviously the publishing of data on the Web, but another is to create the links that allow data to be explored. RDF allows representation of data in such a way that anything referred to in the data can be linked to, and from. If URIs are used to name things, thereby allowing common naming schemes to emerge; one of the most important is the Hypertext Transfer Protocol (HTTP [58]), which affords a straightforward mechanism for people to look up the names. In a properly linked Web of Data, the URI, once looked up, should provide access to useful information about the resource named, as well as useful links out to other data.

Links can be made using various mechanisms, the simplest of which is to use a URI that points to another. For example (taken from [34]), someone might describe some relationships in RDF as follows:

```
<rdf:Description about="#albert"
 <fam:child rdf:Resource="#brian">
  <fam:child rdf:Resource="#carol">
</rdf:Description>
```

This RDF is about three resources given the local identifiers '#albert', '#brian' and '#carol', and might be placed in a file called '<http://example.org/smith>'. The architecture of the Web can use these names to provide a global identifier for the three resources; for instance "http://example.org/smith#albert" refers to #albert, and so on. And now there is a global identifier, links can be made. For instance, a document '<http://example.org/jones>' might contain the following RDF:

```
<rdf:Description about="#denise"
 <fam:child rdf:Resource="#edwin">
  <fam:child rdf:Resource="http://example.org/smith#carol">
</rdf:Description>
```

Here a series of relationships between resources #denise, #edwin and #carol have been asserted, but the datum about #carol makes it possible to link to the data in the other file. Someone following the link dereferences the URI, i.e. decomposes 'http://example.org/smith#carol' into two parts: the part before the '#' which gives the name and location of the file; and '#carol' which is the local identifier in that file. Hence the information about #carol in the first file can be accessed thanks to the link included in the second file. The series of links between different resources can be represented, at least on a small scale, graphically, as in Figure 3. This is the simplest way of linking data, though there are others [34]. And if the URI used for reference is created under a widely-supported system in a community, then the prospects for linking data are that much larger.

One of the main drivers of the SW is the vast quantity of data available in the form of relational databases (RDBs), which often exist in isolation from each other. Each database has its own value, but as argued above, the major source of informational value in the digital age is abundance, the possibilities for serendipitous reuse of data by placing it in fruitful contexts. To that end, a key aim of the SW is to harvest the large amount of data held in RDBs – a much larger quantity than is currently available in the document Web – and to support its amalgamation. The net result will be to

facilitate the treatment of all the data as, in effect, sitting in a single queryable database.

Much of the current Web is supported by larger databases that sit below the level of what can be seen on a webpage; these databases are known as the *deep Web*, as opposed to the *shallow Web* on webpages. So, for instance, when one looks at one's bank statement on the shallow Web, the webpage that the bank's site creates uses a much larger quantity of data about one's bank account than is visible at any one time. Hence an alternative way of looking at this role for the SW is as a method for allowing users to query the whole of the deep Web, rather than simply the information released onto the shallow Web by the current set of implemented applications.

An RDB consists of a series of *tables*, which consist of *rows* or *records* of individual data items. Each record consists of a set of values of *fields* or *attributes*. The records (rows) and fields (columns) together can be conceived as a table or matrix (*m* records and *n* fields give us an $m \times n$ matrix) [49]. So, for instance, each record may represent an individual person, while a particular field may represent zip codes. The value placed in the zip code field of a person's record is therefore the value of that person's zip code. This tabular structure for RDBs can be mapped straightforwardly into an RDF representation. The record can be seen as the subject of an RDF triple, the field name is a link or property type, while the value of that field is the object of the triple. Hence the person who is represented by the record in the example RDB above would be represented by the first item of the triple; the zip code field would be represented by a zip_code property the second item of the triple; and the value of the zip code would be the third item. In that way, each cell of the RDB matrix is represented by an individual RDF triple, and the total set of triples would represent the entire database [33].

Having said that, there are additional factors about RDBs that are harder to capture in the RDF, and there are many open questions about the export of RDB. For instance, it may be that the database is definitive – that is, the institution holding the data has some responsibility for the data. For instance, it may be that the State of Texas holds a database of all Texas vehicle registration numbers; any car not on the database is not, as a matter of definition, registered in Texas. On a smaller scale, some RDBs allow a primary field for a unique identifier for the record, which also holds a significance beyond the particular piece of data. There are various ways of modelling these context-based properties of RDBs, perhaps most likely devolving the representation of such matters to the applications that use the data, via the ontologies, rules or query types that they use. See [33] for a worked example of methods to expose an RDB to the Web.

The process of exposing databases to the Web should not be too prescriptive – the whole point of the Web is as a decentralised collection of linked information, whether in the form of data or documents. The links are entirely democratic, and can be made between any pair of data items, or any pair of documents. This is where the power of the Web's ability to promote serendipitous reuse comes in. Attempting to fix the methods or languages used, or to 'police' the links made, will blur this vision, and create bottlenecks impeding information flow. Indeed, the widespread use of the Web is largely down to its non-prescriptive nature – prescription will simply drive users, who will not want particular information management strategies forced upon them, from the Web.

The result is a particularly untidy situation, unusual in the history of information management. If data, information or knowledge is generated within a single organisation or affiliation (the usual situation for information managers before the growth of the Web), then information systems can trade on a number of simplifying assumptions. The size of such repositories can be assumed to be small or medium, and representation schemes would be planned and homogeneous. The quality of information would be likely to be high, and managers' trust in it correspondingly high. But on the Web scale, these assumptions fail. The amount of data available for query may be extremely high, and represented in highly heterogeneous ways – rarely in the optimal way for the manager's task in hand. Information quality, and trust in that quality, would be very variable. Linking data using common URI schemes will never work perfectly, precisely because there are deliberately no enforcement mechanisms on the Web, and people cannot be forced to use any particular naming convention.

It is primarily for this reason that ontologies have always played a central role in the vision of the SW [56]. It is the ontology that puts the 'semantic' into 'Semantic Web'. Ontologies specify the vocabulary, concepts and relationships of a domain. They must be a rationalisation of current practice, and managed and endorsed by a community. They act as a specification of the terms used in discussion. But they should not be too prescriptive – terms change over time, others are in constant dispute. Ontologies need to develop as a discipline or domain develops. Different areas will have different requirements of ontologies; some sciences will have large, publicly-managed ontologies which act as a public vocabulary standard, while others will make do with small, lightweight ontologies that only define the relationships between a few terms. The ontological requirements of any individual application may be quite small; the question for the application developer is whether to develop a small special-purpose ontology for her own individual purposes, or alternatively whether to reuse a larger, better-known ontology that may overspecify vocabulary for her purposes. Much will depend on the usual practices of her wider community. It should be noted that the SW project does *not* require a single overarching ontology, referring to and prescribing everything.

The ontology is key to being able to deal with heterogeneous datasets as described above; the data, and the terms used in it, must be mapped onto other terms held in common. Once this has been done, then databases can be understood in common terms – and, most crucially, the information they hold amalgamated and processed by machines. It is of course obvious to a human user that if one database of people has a field *ZC*, while another has a field *zip_code*, that there is at least a good chance that the two fields refer to the same attribute of people, viz., the zip codes of their addresses; a computer will merely try, and fail, to match the strings identifying each field. But if the computer is referred to an ontology and given mappings from the terms used in the two databases to the ontology's terms, then it can be told about the equivalence, and accordingly its inference space is opened up (for instance, it could make some inferences from the fact that $ZC(X) = zip\_code(y)$). Machine processing of this heterogeneous, distributed data is made possible by ontologies.

There are, of course, several issues pertaining to the use of ontologies in the SW, some of which will be discussed in greater detail in the 'Controversies' section below. The development of an ontology for an application is a (potentially large) initial cost to an application developer, and it is likely that ontologies, especially well-known ones, will be reused. To that end, searching for ontologies is likely to be a growth area

in the future; there are already specialised search engines, such as Swoogle, dedicated to this task [1, 53]. It may also be that one application might use several ontologies (for instance, an interdisciplinary scientific application may well reuse well-known ontologies from each discipline that it crosses). In that case, mappings *between* the ontologies will be important, and such mappings, as opposed to the ontologies themselves, could become the semantic basis for the application [77].

Building an ontology from scratch is always an option, especially if the application requires only relatively lightweight ontological apparatus [88]; again, special-purpose tools, such as Protégé are already available and well-used in the SW community [1, 89]. Generating an ontology from an RDB can be done semi-automatically, and then mappings (which will be fairly straightforward, given the method of ontology generation) defined between the database and the ontology [104]. The problem is more complex if the aim is to map a legacy database onto an existing ontology. In particular, the mappings between the database and the ontology can be expected to be quite complex, and therefore very expressive languages will be required to describe them, such as the language R_2O [1].

The layered model of the Semantic Web

The Web, as a decentralised construction, cannot be created by fiat or prescription, which would limit its growth and create bottlenecks for information mobility. But to allow the Web of Data to reach fruition at the scale envisaged, several related tasks are required to be performed [100]. As discussed above, the W3C has devoted resources over the last few years to developing formalisms and standards to allow these tasks to be addressed, and the tasks themselves have also been arranged in a series of layers, depicted in a hierarchical diagram. The diagram has evolved with the vision of the SW, but is not dissimilar to its first incarnation, and at the time of writing is seen as in Figure 2.

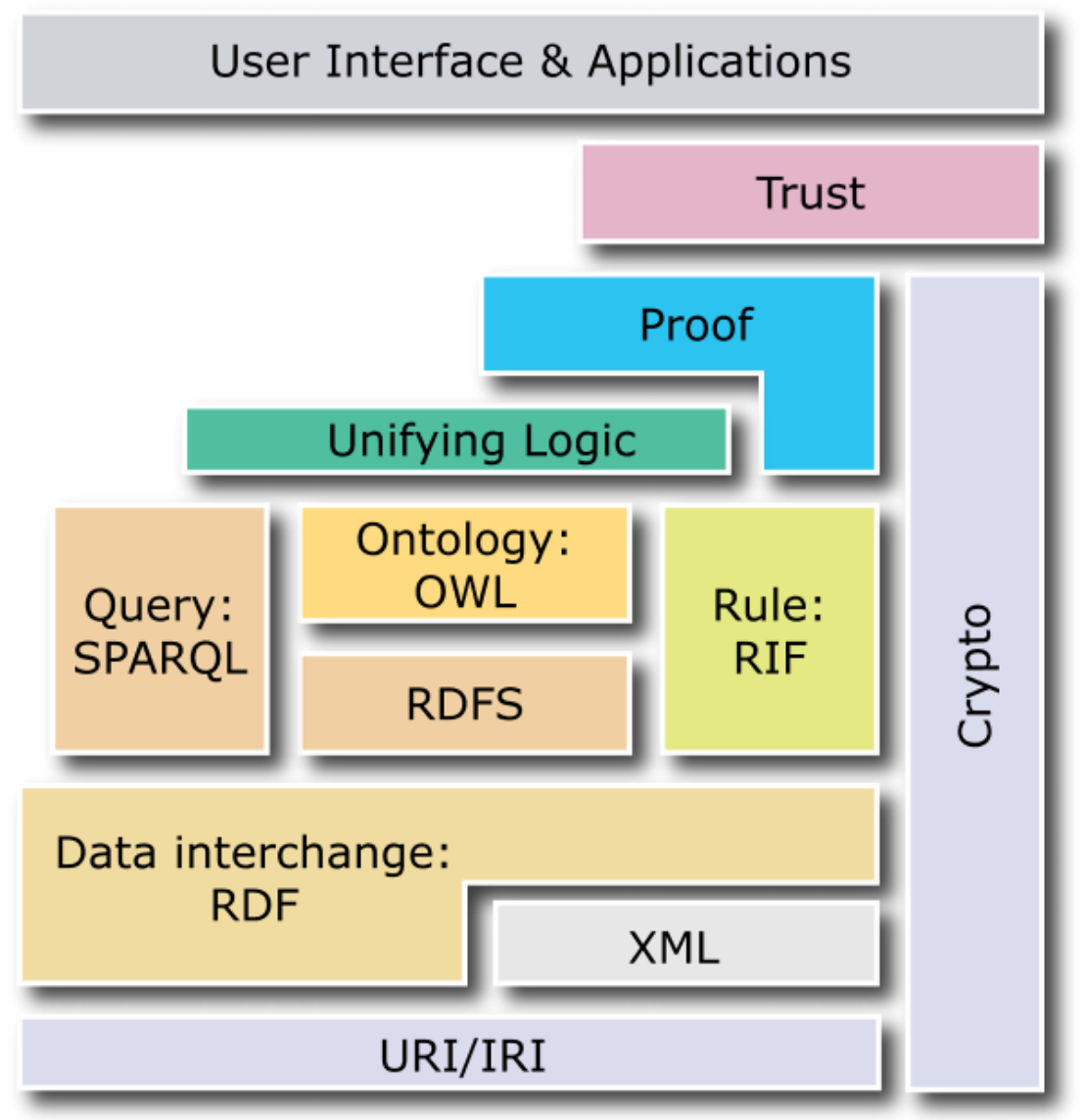

Figure 2: The layered view of the Semantic Web [37]

As noted above, and in Figure 1, the development of these layers has been bottom-up, concentrating on the lower levels before work begins on those further up. The lower layers are now often the subject of W3C recommendations, while work on the upper layers remains generally more theoretical, and contains more open research questions.

At the lowest level we have URIs (and IRIs, Internationalized Resource Identifiers, which are generalisations of URIs allowing non-ASCII characters to be used [55]). URIs identify resources in a global way – in other words, they are interpreted consistently across applications, unlike individual naming conventions, and therefore are central to the vision of a Web of Data [36]. Using a URI to identify a resource (whether that resource be a piece of information, a real-world object, an abstract concept, etc) allows others to use the same identifier to link to the resource, refer to it, or retrieve a representation of it; this shifts the emphasis online from documents to data, and allows direct machine processing of data. If the URI scheme used is the Hypertext Transfer Protocol (http), then that is particularly helpful as http guides the user as to the location of the resource (although there are several other URI schemes, and indeed users can invent their own).

This ability to refer globally is why the export of data from RDBs to the SW should be facilitated, as noted above, by exporting database objects as first-class objects identified with URIs. A number of SW applications have diverged from this vision by not releasing their data onto the Web, but instead archiving them in inaccessible files (at least sometimes because of privacy concerns); Berners-Lee in particular has complained about this tendency [34].

The next layer up from names is that of markup and data interchange – the realm of XML and RDF. The eXtensible Markup Language (XML [43]) is a metalanguage for markup – in other words, a way of supporting communication and data interchange within communities by defining specialised vocabularies – commonly used in a number of sectors.

XML, like the Hypertext Markup Language (HTML) that underpins the current Web, is descended from the Standard Generalized Markup Language (SGML), an international standard for defining system-independent methods of representing information, and so has no conceptual connection to the SW. The main language for data interchange on the SW, on the other hand, RDF, is specially designed for the task, by assigning specific URIs to the fields in its triples. Figure 3 shows an RDF graph of nodes and arcs made up of several triples – each triple consists of two labelled nodes, from which is pointing a labelled directed arc. The two nodes are the first and third elements of the triple; the arc is the second (it points *from* the first element *to* the third). The use of URIs to refer to the properties as well as the objects is an important step to providing semantics – it enables us to reason about and link to relationships as well as objects.

Figure 3: An RDF graph representing Eric Miller [82]

Figure 3 shows four triples, all ‘about’ an individual called Eric Miller, identified by ‘http://www.w3.org/People/EM/contact#me’. If we look at these triples clockwise from the right, the first represents a connection between EM, the property of ‘having the name …’ (which is referred to by the URI ‘http://www.w3.org/2000/10/swap/pim/contact#fullName’), and a character string ‘Eric Miller’. The second links Miller, by the property of having a mailbox, to the value of that property, which is his email address given using the common ‘mailto:’ URI scheme. The third again links Miller with a personal title, the property given as an http URI, and the title as a character string. The fourth triple provides some vocabulary in RDF – it refers to a namespace (an RDF document defining expressive resources which are imported by the RDF graph in Figure 3) which defines some important relations – and in effect says that Miller (the first object in the triple) is an instance of (a relationship which is the second object in the triple) a person (a class which is the third object in the triple).

Based as it is on triples, RDF is simple yet powerful, exploiting the resources of the common subject/predicate/object structure, and its basis in URIs is very important for the SW. It is a minimalist knowledge representation language for the Web – there are some types of knowledge that cannot be represented in RDF, or only represented with difficulty. For instance, a predicate with more than two arguments has to be represented in a somewhat awkward way as a conjunction of two-argument predicates, while statements about hierarchical class relationships, say, need a further formalism. Furthermore, although the graph structure is quite intuitive, the actual syntax of RDF is based on XML (it is called RDF/XML) and, although it is well-suited to machine processing, it is not very easy for the human to read [cf. 26, especially pp.68-69].

The growth in use of RDF has led to the need for special-purpose data stores for holding large quantities of RDF triples (often a set of data will be represented by millions of triples). Such data stores are known as *triplestores*, and need to provide not only storage, but efficient means of reasoning over and retrieving the data that will scale to the large sizes that will be needed. Examples of triplestores include JENA [1], 3store [6, 66] and Oracle 11*g* [15].

Greater expressivity is required than is given in RDF, and to that end there another layer upwards which allows the expression of important information about the vocabularies used to express data. As we can see in Figure 2, the layer here is relatively complex, and contains four boxes, RDFS, Ontology, Rules and Query. These between them provide representation and capabilities that are essential for putting the Web of Data to use.

RDF Schema (RDFS, and sometimes RDF(S) [46]) provides a basic set of tools for producing structured vocabularies that allow different users to agree on particular uses of terms. An extension of RDF, it adds a few modelling primitives with a fixed meaning (such as class, subclass and property relations, and domain and range restriction). It is a basic ontology language that has been adopted fairly widely, and although fairly minimal it can express important constraints on vocabularies.

RDFS is deliberately minimal, and concentrates on expressing subclass and property hierarchies, with various restrictions on these, but the research community, including the Web Ontology Working Group of W3C, identified a number of requirements for greater expressivity for ontologies. As a couple of examples, RDFS allows the stating of subclass relationships, but not, say, that two classes are disjoint; neither does it

allow class cardinality restrictions (e.g. a person has *exactly two* parents). As can be seen in Figure 1, early research efforts into ontology languages led to two leading candidates being developed: DAML (DARPA Agent Markup Language [83]), and OIL (Ontology Interchange Language or Ontology Inference Layer [57]). These two, combined as DAML+OIL [93], became the seed for the W3C ontology language OWL. Unsubstantiated rumour suggests that the fact that it is 'OWL' and not 'WOL' is an arcane joke: in A.A. Milne's *Winnie-the-Pooh* stories, Owl, who is wise and unusually literate for a forest-dweller, spells his name W-O-L. It is more likely that this is a *post hoc* rationalisation of the naming decision.

The needs of ontology languages are great and potentially problematic. There are even problems with the expressivity of RDF: its reification mechanism allows the modeller to make statements about statements – an expressive possibility that can lead to logical problems. RDFS has even more powerful modelling primitives, including 'rdfs:Class', the class of all classes. OWL [84] is a strong language for representing concepts and their relations, and its creators needed to wrestle with the inevitable trade-off between expressivity and efficient reasoning support, with two particular constraints demanded by the Semantic Web. First, there is the strong decentralisation and lack of enforcement mechanisms on the Web, so that people cannot be forced to use a language they do not want to. Creating a language this powerful, for all purposes, may well have resulted in it not being used at all. Those who need great expressivity might be inclined to use their own favourite non-standard language, while those who want efficient reasoning might prefer to revert to RDFS. The second constraint is that the layered view of the SW (Figure 2) makes it desirable that OWL should be an extension of RDFS – it should use the RDF interpretation of classes and properties and add primitives to provide richer expressivity. However, this cannot be the basis for OWL, because the addition of powerful reasoning to the expressive power of RDFS (to define such items as the class of all classes) would result in a language very hard to control

To pick their way between these various pitfalls, the developers of OWL created three separate languages [84]. OWL Full is the complete language, a full set of OWL primitives, which can be combined with RDF and RDFS in arbitrary ways. This includes the possibility that an OWL Full ontology could augment or alter the meaning of a pre-defined RDF, RDFS or OWL term (for instance, one could put a cardinality constraint on the size of the class of all classes, thereby limiting the number of possible classes that could be constructed). OWL Full is compatible with RDF, so that any legal RDF document is an OWL Full document. The downside of all this expressivity is that the language is undecidable, which rules out the possibility of complete reasoning support.

OWL DL is intended to be open to computational support, and is a sublanguage of OWL Full (so that any legal OWL DL ontology is a legal OWL Full ontology). OWL DL is so named because it is based on *description logic*, a type of knowledge representation language used to describe the knowledge definitive of a particular application domain [28], and is designed to be complete and decidable (in particular, application of OWL's constructors to each other is restricted). This does mean that full compatibility with RDF is lost: although every legal OWL DL document is a legal RDF document, the inverse is not true.

OWL Lite is a very lightweight language to support users requiring a classification hierarchy with simple constraints. Reasoning support should be relatively efficient,

and it is intended to provide a straightforward migration route for bringing thesauri and taxonomies to the SW. A legal OWL Lite ontology is also a legal OWL DL ontology, but the cost of ease of reasoning is a lack of expressivity, so for instance enumerated classes, statements of disjointness and assignment of arbitrary cardinality constraints (i.e. restrictions of class cardinality to any number other than 0 and 1) are disallowed.

Even though the relationship between OWL and RDF is complex, its roots in RDF allow OWL to exploit RDF's linking capabilities to allow ontologies to be distributed across systems. When constructing an ontology in OWL, the developer can refer to terms in other ontologies, which then encourages the sharing of terminology across distributed data sources. Sharing ontologies is not always sufficient when it comes to data sharing – an organisation may find that nearly all of an imported ontology is adequate, but it needs extra identifiers and descriptions, and in such a case it should be allowed to add them, rather than build a new ontology from scratch [69]. This ability to assemble distributed ontologies is central to the SW vision.

Expressing ontological relations is a central, perhaps the most controversial (see below), part of the SW vision, but having achieved a representation of the domain with semantics, one still needs to make inferences. OWL has some inferential support, such as subsumption and classification, but there are several inferential methods that will be required on the SW. Hence, work is currently ongoing on the Rule Interchange Format (RIF), which is intended to allow a variety of rule-based formalisms, ranging from Horn-clause logics, higher order logics and production systems, to be used [40].Various insights from Artificial Intelligence (AI) have also been adapted for use for the SW for various purposes, including temporal (time-based) logic, causal logic and probabilistic logics [100, 74, 101, 29].

And given the domain description at its desired level of expressivity and a means of making inferences, then the next important function at this level is the ability to query the data. Once more a special-purpose language is being developed by the W3C, which at the time of writing is very close to completion and achieving recommendation status. SPARQL works in effect by constructing a graph of RDF-like triples that may contain variables, which is then matched against the RDF graph to be queried; the query is successful if there is a subgraph of the RDF graph which matches the query when RDF terms are substituted for its variables [95].

Sitting on top of these layers are further layers with a unifying logic, proof systems and trust systems. As can be seen from Figure 1, these upper layers remain topics of exploratory research.

Trust is perhaps key to widespread application of the SW. If information is being drawn from heterogeneous sources, then it is important that users are able to trust such sources if they are to act on the inferences that result. Trust will of course depend on the criticality of the inferences – trust entails risk, and a risk-averse user will naturally trust fewer sources [91, 41]. Measuring trust, however, is a complex problem [61]. A key parameter is that of the provenance of data, a statement of the conditions under which data were produced (including statements about the methods of production and the organisation that carried them out). Methods are appearing to describe provenance [62], but more needs to be discovered about how information spreads across the Web, and therefore how it can be tracked and understood [37].

Related issues include respect for intellectual property, and the privacy of data subjects. In each case the reasoning abilities of the SW can be of value, and initiatives are currently under way to try to exploit them [92]. Protocols that allow users to express their own privacy preferences, and to enable those who wish to reuse information to reason about those preferences, are being created under the programme of research into the Policy Aware Web [107]. Creative Commons is an initiative for representing copyright policies and preferences based on RDF to promote reuse where possible (current standard copyright assumptions are deliberately restrictive with respect to reuse) [1]. Cryptography protocols to protect information and privacy will also play an important role at all levels, as shown in Figure 2.

Applications

The top layer of Figure 2 is that of a user interface and applications. This recognises the fact that if the SW cannot be used easily, and integrated into people's workflows in order to add value to their informational transactions, then it will not attract a large user base, without which the network effects already seen in the development of the World Wide Web will not transpire. Network effects are those positive benefits that increase in certain communication systems faster than its user base expands. In the same way as a telephone system is of limited value to a handful of people and enormous value to a large number, a few people exposing data to the SW is unlikely to make much of a difference, whereas if scalable SW technologies were applied to something like the quantity of data to be found currently in the Deep Web, the gains would be immense.

One not entirely frivolous way of expressing the need for the top layer of the SW is to say that its user base needs to grow quickly, and what is needed is a 'killer app', in other words an application that will meet a felt need and create a perception of the technology as 'essential'. Less ambitiously, the SW's spread depends not only on having an impressive set of formalisms, but also the tools to use the linked data [25].

Bootstrapping

One particular user issue is the importance of bootstrapping content for the SW. Even if RDF began to be published routinely, the amount of legacy content on the Web would dwarf new data for some time, and to make this legacy accessible to SW technology some automation of the process of creating RDF from other formats is required. CS AKTive Space [97], discussed in more detail below, amassed a large quantity of information about the state of computer science research in the United Kingdom through a relatively laborious process of harvesting information from the webpages of computer science departments in British universities without necessarily acting as a source, and using natural language processing and an ontology to interpret the data. The application was very successful, but the researchers were SW researchers, and the process is likely to be too onerous to be repeated on a large scale by non-experts. Assumptions can be made about webpage structure (for example, about regular layouts generated from a database by an individual website), and tools have been developed to exploit them [81].

An important development in this field is GRDDL (Gleaning Resource Descriptions from Dialects of Languages) which became a W3C recommendation in September 2007, allows the extraction of RDF from XML and XHTML (a further markup language) documents using transformations expressed in XSLT, an extensible stylesheet language based on XML. It is hoped that such extraction could allow

bootstrapping of some of the hoped-for SW network effects, given the amount of XML and XHTML data in the Deep Web [51].

Annotating documents and data with metadata about their content, provenance and other useful dimensions (even including the emotional dimension to content – [99]) is also important for the effort to bring more content into the range of SW technologies [63]. Multimedia are a particular focus for research into annotation [105]. Manual annotation is a great burden for information holders, and a major initial cost for the SW, so methods of automating annotation have been investigated by a number of research teams in order to increase the quantity of annotated data available without excessive expenditure of resources [63, 64, 106].

In addition, as a large quantity of the Web is actually written in natural language, some have seen a role for natural language processing (NLP), and information extraction (IE), for analysing this text statistically. So large is the Web's store of written language (two thousand billion words) that it can function as a corpus which dwarfs the most ambitious attempts of dedicated corpus builders in computational linguistics of only a few years ago [78]. And given this, and the need for automating or semi-automating annotation, NLP techniques, augmented by ontologies and training with humans, can be used to extracted machine-readable structured information from plain text [42, 47, 48, 75]. There have also been attempts to build ontologies using NLP techniques, another of the major anticipated bottlenecks for the SW [44]. See also the section 'Commercial and non-academic applications' on p.20 below for some examples of SW applications using NLP.

Application areas

Predicting which particular applications will succeed is unscientific and usually inaccurate. In fact, as it is the rich information contexts that the web of linked data provides that will increase the value of individual pieces of data, one way in which such growth can be encouraged is to focus on small communities with pressing information-processing requirements, and various more-or-less common goals; such communities can be the 'killer apps', or, more accurately, the early adopters of the technology, exactly as the high energy physics discipline played a vital role in the development of the WWW [cf. e.g. 19]. A series of case studies and use cases is maintained at [20].

The most promising of these communities is *e-science*, the data-driven, computationally-intensive pursuit of science in highly distributed computational environments [71]. Large quantities of data are created by analyses and experiments in disciplines such as particle physics, meteorology and the life sciences. Furthermore, in many contexts, different communities of scientists will come together to perform interdisciplinary work, so that data from various fields (e.g. genomics, clinical drug trials and epidemiology) varying not only in vocabulary, but also in the scale of description, need to be integrated. Many scientific disciplines have created large-scale and robust ontologies for this and other purposes. The most well-known of these is the Gene Ontology, a controlled vocabulary to describe gene and gene product attributes in organisms, and related vocabularies developed by Open Biomedical Ontologies. Others include the Protein Ontology, the Cell Cycle Ontology, MeSH (Medical Subject Headings, used to index life science publications), SNOMED (Systematized Nomenclature of Medicine) and AGROVOC (agriculture, forestry, fisheries and food). For more examples and references see [100].

E-government is another important application area, where heterogeneous information of varying quality is deployed widely. Government information varies in provenance, confidentiality and "shelf life" (some information will be good for decades or even centuries, while other information can be out of date within hours), while it can also have been created by various levels of government (national/federal, regional, state, city, parish). Privacy and security are also obviously important factors in this space. Integrating government information in a timely way is clearly an important challenge (see for instance a pilot study for the United Kingdom's Office of Public Sector Information, exploring the use of SW technologies for disseminating, sharing and reusing data held in the public sector [24]).

Academic applications

Many applications for the SW have been developed with the specific purpose of bringing the SW to maturity. These are often written up in conferences such as the regular World Wide Web Conferences, the International Semantic Web Conferences (ISWC), the European Semantic Web Conferences (ESWC), as well as several one-off conferences and workshops, and can be found in the proceedings (usually online) of any of these. See also the *Journal of Web Semantics* [21].

One initiative of interest here is the Semantic Web Challenge [10], which runs annually alongside the ISWC. This is a good-natured competition to find applications that show SW technology in the best light and which can act as benchmarks for the research community. These applications, therefore, are to some extent an objective list of applications through the years that use semantic technologies to solve real-world problems involving heterogeneous real-world data. The winners of the SW Challenges from its inception, 2003, to the time of writing, 2007, are as follows.

2003: CS AKTive Space (University of Southampton) is an application to explore the UK Computer Science Research domain across multiple dimensions for multiple stakeholders, allowing the tracking of the activities of all agents from funding agencies to individual researchers, using information harvested from the Web, and mediated through an ontology [97].

2004: Flink (Vrije Universiteit Amsterdam) is a 'Who's Who' of the SW which allows the interrogation of information gathered automatically from Web-accessible resources about researchers who have participated in ISWC conferences [85].

2005: CONFOTO (appmosphere web applications, Germany) is a browsing and annotation service for conference photographs [87].

2006: MultimediaN E-Culture Demonstrator (Vrije Universiteit Amsterdam, Centre for Mathematics and Computer Science, Universiteit van Amsterdam, Digital Heritage Netherlands and Technical University of Eindhoven) searches, navigates and annotates media collections interactively, using digital representations of items from the collections of several well-known museums and art repositories [98].

2007: Revyu.com (Open University) is a reviewing and rating site specifically designed for the SW, allowing reviews of any kind of resource, content or event to be integrated and interlinked with data from other sources (in particular, other reviews, which proliferate on the Web) [68].

A typical SW application will generate a new ontologies for its application domain (e.g. art, as with MultimediaN or computer science, as with CS AKTive Space), and use it to interrogate large stores of data, whether legacy data or freshly harvested. This

strand of research is tending to confirm the hypothesis that ontologies have an important role in mediating the integration of data from heterogeneous sources.

Commercial and non-academic applications

SW applications are generally presented using custom-built interfaces. This suggests a very important area for future research, the development of scalable visualisers capable of navigating the graph of connected information expressed in RDF. As can be seen, the importance of applications and user interfaces was made clear in the layered SW diagram (Figure 2). However, we shouldn't expect to 'see' the SW in a special browser, in the way that we can see the Web of Documents through browsers such as Internet Explorer, Netscape or Mozilla Firefox. Rather, SW technologies, facilitating the exploration of data, may well work at the back end of websites to improve the user experience. Examples of such sites pointed to by the W3C include Sun's white paper collection site [17], Nokia's developers' discussion forum [10], Oracle's virtual press room [4], and Harper's online magazine [12].

There is an increasing number of applications supporting deeper querying of linked data. The DBpedia [27] is based on the collaborative encyclopaedia Wikipedia created by volunteers, and is intended to extract structured information from Wikipedia allowing much more sophisticated querying. Sample queries given on the DPpedia website include a list of people influenced by Friedrich Nietzsche, and the set of images of American guitarists. DBpedia uses RDF, and is also interlinked with other data sources on the Web. When accessed in late 2007, the DBpedia dataset contained 103 million RDF triples. Other examples of linked data applications include the DBLP bibliography of scientific papers [22], and the GeoNames database which represents descriptions of millions of geographical features in RDF [11].

As well as existing organisations using semantic technologies to improve user experience, and applications exploiting linked data, commercial firms are beginning to appear whose business model is based on the possibilities of the SW. Garlik [23] aims to provide individual consumers with more power over their digital data. It reviews what is held about people, harvesting data from the open Web, and represents this in a people-centric structure. Natural Language Processing is used to find occurrences of people's names, sensitive information, and relations to other individuals and organisations.[1] Twine [18] aims to facilitate knowledge and information sharing, and to organise that information using various SW technologies (also, like Garlik, using NLP). Twine's developer Nova Spivack coined the term 'knowledge networking' to describe the sharing process, analogous to the Web 2.0 idea of 'social networking'.

Controversies

The SW has been controversial during its history, with several commentators arguing that it is based upon unrealistic expectations, or repeats the mistakes of other initiatives. The arguments against the SW have tended to appear more in the blogosphere rather than the academic world, perhaps because people in the SW world are genuinely enthusiasts while those without confidence in the SW project are doing other things. The pro-SW website GetSemantic supports a wiki page of arguments against the SW, with references and responses [2]. In this section, we will examine three of the most prominent arguments raised against the SW.

[1] Declaration of interest: Wendy Hall is Chair of the Garlik Advisory Board.

The Semantic Web repeats the mistakes of "Good Old-Fashioned Artificial Intelligence"

It has been argued that the SW is basically a throwback to the project to programme machine intelligence [76] which was jokingly christened by John Haugeland 'GOFAI' (Good Old-Fashioned AI). GOFAI proved impossible: so much of human intelligence is implicit, context-dependent and situated that writing down everything a computer needs to know to produce output that exhibits human-like intelligence is out of the question [67].

One attempt to work around this problem is the Cyc project, set up in 1984, which aims to produce a gigantic ontology that will encode all common-sense knowledge, in order to support human-like reasoning by machines (i.e. GOFAI) [80]. The project has always aroused controversy, but it is fair to say that over two decades later, GOFAI is no nearer. The implicit nature of common-sense knowledge arguably makes it impossible to write it all down.

Many commentators have argued that the SW is basically a re-creation of the (misconceived) GOFAI idea, that the aim is to create machine intelligence over the Web, to allow machines to reason about Web content in such a way as to exhibit intelligence [76]. This, however, is a misconception, possibly abetted by the strong focus in the 2001 *Scientific American* article on an agent-based vision of the SW [39], although co-author of that article James Hendler has stated very firmly that he believes that the article was radically misinterpreted, and that no-one "can … say we're advocates of the big AI vision when we explicitly make it clear we're pushing for something else" [70]. The *Scientific American* article states that "Traditional knowledge-representation systems typically have been centralized, requiring everyone to share exactly the same definition of common concepts such as 'parent' or 'vehicle.' But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable" [39].

The SW is not GOFAI reheated, but rather an attempt to facilitate sharing of, and context-based machine reasoning over, content (and therefore the provision of machine-readable data on the Web). The aim is not to bring a single ontology, such as Cyc, to bear on all problems (implicitly defining or anticipating all problems and points of view), but to allow data to be interrogated in ways that were not anticipated by their creators. Different ontologies will be appropriate for different purposes; composite ontologies can be assembled from distributed parts [77, 84]. and it is frequently very basic ontologies (defining simple terms such as 'customer', 'account number' or 'account balance') that add most value to content. In this respect, the situation in the SW simply mirrors offline life where people from different communities and disciplines can and do interact without making any kind of common *global* ontological commitment [100, 37, 35]. The engineering challenge, as Berners-Lee et al argue, is to allow independent consistent data systems to be connected locally without requiring global consistency [38].

Yorick Wilks, accepting that the SW is not an attempt to recreate GOFAI, argues that this is both a gain and a loss: a gain because the knowledge representation structures the SW proposes are computationally tractable, as opposed to the various GOFAI formalisms; a loss because DAML+OIL (and presumably by extension OWL) is less sophisticated than those formalisms, and may not have the representational power for the complexity of the world, whether common-sense or scientific [108]. Equally, as both Wilks and Berners-Lee point out, many in the SW world began their research

careers in artificial intelligence, as Shadbolt et al argue that "it will draw on some key insights, tools and techniques derived from 50 years of AI research' [100].

Ontologies

Ontologies, as we have seen, are vital for the SW vision of a Web of Data, but are perceived by many as expensive to develop and hard to maintain. The ideal conceptual apparatus is relative to the task in hand, and different ontologies are appropriate for different tasks. Classifications are also made relative to some background assumptions, and impose those assumptions onto the resulting ontology. To that extent, the expensive development of ontologies reflects the world view of the ontology builders, not necessarily the users. They are top-down and authoritarian, and therefore opposed to the Web ethos of decentralisation and open conversation. They are fixed in advance, and so they don't work very well to represent knowledge in dynamic, situated contexts. [94] argues, for instance, that ontologies do not capture the situated processes of scientific research, the social construction of knowledge or the emergence and evolution of understanding over time, and presents an alternative way of representing this knowledge. [103] implicitly endorses this view, showing how there are issues in biology that OWL DL is not well-equipped to handle.

Other papers have made the point that some types of knowledge are more naturally modelled in ontologies than others, and, while not opposing the use of ontologies, warn against too strong a reliance on them for knowledge representation. [60] argues that ontologies cannot be too ambitious, and attempts to reify the context of an ontology (i.e. to provide context-independent accounts of knowledge) will be undermined by knowledge's situated nature. [90] argues that the social context of knowledge requires application builders to be maximally receptive to diverse types of heterogeneous reasoning, which might use knowledge that is hard to capture in hierarchical structures. See also [45] for a series of short essays debating this point.

A related critical point is that the Web as a decentralised, linked information structure must reflect the pragmatic needs of its large, heterogeneous user base which includes very many people who are naïve in their understanding of computing issues. The infrastructure has to be usable widely, which argues for simplicity. The rich linking structure of the Web of Documents, combined with statistically-based search engines such as Google, is much more responsive to the needs of unsophisticated users. The SW, in contrast, demands new information representation, markup and publishing practices, and corporations and information owners need to invest in new technologies. Not only that, but current statistical methods will scale up as the number of users and interactions grows, whereas logic-based methods such as those advocated by the SW, on the other hand, scale less well [cf. e.g. 110].

The dispute has been fuelled by the flowering since 2005 or so of the so-called 'Web 2.0' paradigm (of systems, communities and services facilitating collaboration and information-sharing among users). In particular, it has been argued that the meaningful structures that emerge when sufficiently large numbers of users 'tag' content with key words, structures which have been called 'folksonomies', resulting in a structure of connections and classifications emerging without central control, 'really' express the assumptions of the users, and furthermore in such a way as to respect their familiar patterns of communication and workflow. Meanwhile, ontologies 'really' express the needs of the ontology developers and their sponsors [102].

However, folksonomies are much less expressive than ontologies; they are basically variants on keyword searches. A tag 'SF' may refer to science fiction or San Francisco, even if we make the unrealistic assumption of a monoglot English user community. In a multilingual environment such as the Web, further ambiguity is possible – for instance, 'SF' might refer to the Swiss television station Schweizer Fernsehen. Furthermore, the semantics of Web 2.0 are relatively shallow, with few links and very sparse hierarchies.

When a community is large enough and the benefits clear enough to provide incentives to work together, then a large-scale ontology building and maintenance programme is justified. It is true that large fixed costs will tend to skew the effort involved towards authorities who may be unrepresentative [90], but Shadbolt et al argue explicitly that "the ontologies that will furnish the Semantics for the Semantic Web must be developed, managed, and endorsed by committed practice communities. Whether the subject is meteorology or bank transactions, proteins or engine parts, we need concept definitions we can use" [100].

It is of course an undecided question as to whether this community involvement will transpire, but in a recent note, Berners-Lee argues that such conditions will be perhaps more frequently encountered than sceptics believe. On the broad assumption that the size of an ontology-building team increases on the order of the log of the size of the ontology's user community, and that the resources needed to build an ontology increase on the order of the square of community size, the cost per individual of ontology building will diminish rapidly as community size increases. These assumptions are explicitly intended to be indicative rather than realistic [35].

More to the point, not all ontologies need be of great size and expressive depth. It is certainly not the case that the SW requires a single ontology of all discourse on the model of Cyc. Such an ontology, even if possible, would not scale, and in a decentralised structure like the Web its use could not be enforced. Even in complex scientific domains, [73] argues, using a case study from the field of medical informatics, that ontologies should be firmly based on work practices in the domain. In more mundane applications, we should expect a lot of use of small-scale, *shallow* ontologies defining just a few terms that nevertheless are widely applicable [35].

For example, the machine-readable Friend-of-a-Friend (FOAF) ontology is intended to describe people, their activities and their relations to other people. It is not complex, and publishing a FOAF profile is a fairly simple matter for which there are dedicated tools [14]. The resulting network of people has become very large indeed. A survey performed in 2004 discovered over 1.5 million documents using the FOAF ontology [54].

In any case, ontologies and folksonomies serve different purposes. Folksonomies are based on word tags, whereas the basis for ontology reference is via a URI. One of the main aims of ontology definition is to *remove* ambiguity – not globally, for this may well be impossible, but rather within the particular context of the application. Folksonomies will necessarily inherit the ambiguity from the natural language upon which they are based. Nevertheless, a strong possibility that has been considered is to use cheaply-gathered folksonomies as starting points for ontology development, gradually morphing the Web 2.0 structures into something with greater precision and less ambiguity [86, 72].

Symbol grounding

An important aspect of the SW is that URIs must be interpreted consistently. However, terms and symbols are highly variable in their definitions and use through time and space. The SW project will be boosted by processes whereby URIs are given to objects by communities and individuals, endorsed by the user community, who ensure consistency. Responsible URI 'ownership' is critical to the smooth functioning of the SW [100].

But the process of ensuring a fixed and known link between a symbol and its referent, which has been called *symbol grounding*, is at best hard [65], and at worst impossible [109]. Meanings do not stay fixed, but alter, often imperceptibly. They are delineated not only by logical definitions in terms of necessary and sufficient conditions, but also by procedures, technologies and instrumentation, and alter subtly as practice alters.

Any attempt to fix the reference of URIs is a special case of symbol grounding, and is consequently hard to do globally. Attempting to resist the alteration in community practices and norms, and reformulation of meanings of terms, would be doomed. This is understood by leading developers of the SW, who agree that "communities and practice will change norms, conceptualizations, and terminologies in complex and sociologically subtle ways. We shouldn't be surprised or attempt to resist these reformulations" [100]. But there is an important issue, as the same authors concede. "The issue for a Semantic Web built [in a community-driven way] is to know when parts need revision" [100].

Yorick Wilks has argued that Natural Language Processing techniques are essential for grounding the SW, because of the preponderance of text-based content on the Web. NLP is central as the procedural bridge from texts to knowledge representation, usually via automatic information extraction [108]. Berners-Lee has argued in response to Wilks, at a Web Science Workshop in 2005 that the SW was necessarily based on logic and firm definitions (even if those definitions were imperfect, or highly situated and task-relative), not words, use patterns or statistics. Though meanings are not fully stable, they can be stable *enough* relative to individual applications and in particular contexts to allow the SW approach to work [9]. In the case of large-scale, deep ontologies describing sciences, that perhaps will be where the SW is likely to add most value, the Berners-Lee view is reminiscent of that of Hilary Putnam that scientists are 'guardians' of meaning, who determine the 'true' referent of a word like 'water' [96]. But Berners-Lee agrees that ontologies will need to evolve – some quite quickly, and that such meanings cannot be fixed irrevocably; nevertheless, for the purposes of particular applications, this is unlikely to be a problem in practice [100].

Future directions

The SW is a work in progress, though Shadbolt et al argue that the need for shared semantics and a Web of Data have increased, and furthermore that the SW is "attainable" [100]. This final section will sketch some of the anticipated directions of future SW work.

Standards

The most obvious future direction is to continue the research as planned. The development of the SW has been conceived as a tide rolling over a beach, covering some areas fully, enveloping other areas more slowly [Figure 1]. As has been noted, the upper layers of the SW, looking at trust, logic and proof, are relatively

underdeveloped, and are the focus for exploratory research at the cutting edge. The lower layers of the SW are in place and deployed widely. The middle layers are more or less in place; OWL is complete, while SPARQL and RIF are progressing, and should both become W3C recommendations in the fullness of time.

The Semantic Grid

Grid computing is a type of distributed computing designed to apply computational power from a number of different distributed, complete computers working in parallel, and in cooperation, on a single problem. For some extremely data-heavy problems requiring a lot of computation (particularly in e-science), grid computing is an important time-saving solution. Particular issues in grid computing include the problems of coordinated resource sharing, distributed problem-solving and the creation of 'virtual organisations' to pool data and share outcomes. The SW, of course, is another distributed computing paradigm where data sharing is a key issue – with the SW, a Web of Data, sharing is the whole point. A third distributed paradigm – software agents – is also a relevant factor.

This synergy has led to a research strand to apply semantic technologies to the problems of grid computing, adding meaning via ontologies and RDF metadata annotations to the grid. Information and services for the grid are thereby given well-defined meaning, which enables the interaction between humans and computers to be better coordinated. In particular, all the components, services and resources are adequately described for machine processing. The use of semantics to describe grid resources is known as the *Semantic Grid*, and research is ongoing [16, 71, 59, 52].

The policy-aware Web

As is clear in Figure 2, trusted systems are very important to the development of the SW. There are two reasons for this. First, if someone is reasoning with heterogeneous data harvested from the Web, then they will need to trust the data they have harvested and are using. As noted above, research is ongoing into methods for specifying the provenance of such data [62]. The second reason is that people will not release their data if they thought it would be misused; the importance of data privacy in our digital age is easily underestimated [92]. The *policy-aware Web* is an initiative designed to rectify this problem.

The assumption behind the policy-aware Web is that inflexible and simplistic security systems and access control for the decentralised environment of the Web has hampered its development. Insufficiently sophisticated controls have made people reluctant to share data, particularly with other parties with which they do not have pre-existing information-sharing policies. Furthermore, the Web of Documents is rather coarse-grained for detailed security: the security decision to be made is to grant access to an entire website or page, or not, because policy control mechanisms for access at a finer-grained level aren't available. Thus, despite increasing amounts of useful information residing on the Web in a machine-retrieval form, reluctance to share that information remains.

The aim of policy-aware Web technology is to provide for the publication of access policies in a way that allows significant transparency for sharing among partners without requiring pre-agreement. In addition, greater control over information release can be placed in the hands of the information owner, allowing discretionary (rather than mandatory) access control to flourish. Policies would be another kind of

metadata attached to information, and those wishing to use that information would be able to reason about them. For instance, one should be able to specify that the information can only be used by the agent gaining access, and that that agent should not pass the information on. Or it may be specified that the information should be deleted after a certain period of time. Or if it is to be used in a certain manner, then data should be anonymised.

Enforcement of these policies is another matter, but at present the research effort is focused on how to express such policies, and on creating theorem provers to reason about them. The result should be a much more fine-grained security picture, with greater transparency and accountability of information use [107].

Web Science

Although since its inception the Web has revolutionised communication, collaboration and education (particularly within science), relatively little is known about the way it develops. There is a growing feeling among researchers across a number of disciplines that a clear research agenda aimed at understanding the current, evolving and potential Web is needed to assure its continued growth. Such researchers want to model the Web, understand the architectural principles that have provided for its growth, and be as sure as possible that it supports the basic social values of trustworthiness, privacy, and respect for social boundaries, and their solution is to chart out a research agenda that targets the Web as a primary focus of attention [38, 37].

This agenda has been dubbed *Web Science*, a combination of analysis of the Web and its dynamics, and synthesis of new languages and protocols. The Web is an engineered space created via formally specified languages but, as humans are the creators of Web pages and links between them, their interactions form emergent patterns in the Web at a macroscopic scale. These human interactions are, in turn, governed by social conventions and laws. Web Science is, therefore, inherently interdisciplinary; its goal is to both understand the growth of the Web and to create approaches allowing new powerful and more beneficial patterns to occur.

Such a research area does not yet exist in a coherent form. Within computer science Web-related research has largely focused on information retrieval algorithms and the algorithms for the routing of information through the underlying Internet. Outside of computing, researchers grow ever more dependent on the Web, but there is no concerted agenda for exploring emerging trends on the Web nor are those outside computer science fully engaged with the emerging Web research community to focus more specifically on the needs of science and of society as a whole, while preserving the essential invariants of the Web experience: decentralisation to avoid social and technical bottlenecks, openness to the reuse of information in unexpected ways, and freedom and equality of information as it passes across the Web.

Despite excitement about the SW, the majority of the world's data is locked in large data stores and is not published as an open web of inter-referring resources. As a result, the reuse of information has been limited. Substantial research challenges arise in changing this situation. We have already discussed the need for policy controls, and for tools to allow scientists to exploit data when it emerges. But on top of that, releasing data is both a technical and a social problem, and understanding how to free data to the SW is a matter of understanding society in relation to the Web (in social, legal and economic terms) and the Web in relation to society. This is the foundation

of the emerging Web Science agenda which it is hoped will inform the development of the SW [100]. The recent foundation of the Web Science Research Initiative (WSRI [8]), a joint venture between the Massachusetts Institute of Technology and the University of Southampton, is intended to drive the agenda on, acting as a focus (e.g. advising in particular on curricula to support it).

Bibliography

Primary literature

1. http://creativecommons.org/about/ (accessed December 2007).
2. http://getsemantic.com/wiki/Arguments_against_the_Semantic_Web (accessed December 2007).
3. http://jena.sourceforge.net/ (accessed December 2007).
4. http://pressroom.oracle.com/ (accessed December 2007).
5. http://protege.stanford.edu/ (accessed December 2007).
6. http://sourceforge.net/projects/threestore (accessed December 2007).
7. http://swoogle.umbc.edu/ (accessed December 2007).
8. http://webscience.org/ (accessed December 2007).
9. http://www.cs.umd.edu/~hendler/2005/WebSciReport.pdf, 2005 (accessed December 2007).
10. http://www.forum.nokia.com/ (accessed December 2007).
11. http://www.geonames.org/ (accessed December 2007).
12. http://www.harpers.org/ (accessed December 2007).
13. http://www.informatik.uni-bremen.de/agki/www/swc/index.html (accessed December 2007).
14. http://www.ldodds.com/foaf/foaf-a-matic (accessed December 2007).
15. http://www.oracle.com/technology/tech/semantic_technologies/index.html (accessed December 2007).
16. http://www.semanticgrid.org/ (accessed December 2007).
17. http://www.sun.com/servers/wp.jsp (accessed December 2007).
18. http://www.twine.com/ (accessed December 2007).
19. http://www.w3.org/2001/sw/SW-FAQ (accessed December 2007).
20. http://www.w3.org/2001/sw/sweo/public/UseCases/ (accessed December 2007).
21. http://www.websemanticsjournal.org/ (accessed December 2007).
22. http://www4.wiwiss.fu-berlin.de/dblp/ (accessed December 2007).
23. https://www.garlik.com/index.php (accessed December 2007).
24. Alani, H., Dupplaw, D., Sheridan, J., O'Hara, K., Darlington, J., Shadbolt, N.; Tullo, C. (2007) Unlocking the potential of public sector information with Semantic Web technology. In Proceedings of the 6th international Semantic

Web conference 2007, Busan, South Korea, http://iswc2007.semanticweb.org/papers/701.pdf (accessed December 2007).

25. Alani, H., Kalfoglou, Y., O'Hara, K., Shadbolt, N. (2005) Towards a killer app for the Semantic Web. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (Eds.) The Semantic Web, proceedings of the international Semantic Web conference 2005, Hiroshima, Japan. Springer, Berlin, 829-843.
26. Antoniou, G., van Harmelen, F. (2004) A Semantic Web primer. MIT Press, Cambridge, MA.
27. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. (2007) DBpedia: a nucleus for a Web of open data. In Proceedings of the 6th international Semantic Web conference 2007, Busan, South Korea. http://iswc2007.semanticweb.org/papers/715.pdf (accessed December 2007).
28. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (Eds.) (2003) The description logic handbook: theory, implementation and applications. Cambridge University Press, Cambridge.
29. Baclawski, K., Niu, T. (2005) Ontologies for bioinformatics. MIT Press, Cambridge, MA.
30. Barrasa, J., Corcho, O., Gómez-Pérez, A. (2004) R_2O, an extensible and semantically based database-to-ontology mapping language. 2nd Workshop on Semantic Web and Databases (SWDB2004), Toronto, http://www.cs.man.ac.uk/~ocorcho/documents/SWDB2004_BarrasaEtAl.pdf (accessed December 2007).
31. Baumol, W.J. (2002) The free-market innovation machine: analyzing the growth miracle of capitalism. Princeton University Press, Princeton.
32. Berners-Lee, T. (1994) Plenary at WWW Geneva 94, http://www.w3.org/Talks/WWW94Tim/ (accessed December 2007).
33. Berners-Lee, T. (1998) Relational databases on the Semantic Web. http://www.w3.org/DesignIssues/RDB-RDF.html (accessed December 2007).
34. Berners-Lee, T. (2006/2007) Linked data. http://www.w3.org/DesignIssues/LinkedData.html (accessed December 2007).
35. Berners-Lee, T. (2007) The fractal nature of the Web. http://www.w3.org/DesignIssues/Fractal.html (accessed December 2007).
36. Berners-Lee, T., Fielding, R., Masinter, L. (2005) Uniform Resource Identifier (URI): generic syntax. http://gbiv.com/protocols/uri/rfc/rfc3986.html (accessed December 2007).
37. Berners-Lee, T., Hall, W., Hendler, J.A., O'Hara, K., Shadbolt, N., Weitzner, D.J. (2006) A framework for Web Science. Foundations and Trends in Web Science 1(1), 1-134, http://www.nowpublishers.com/product.aspx?product=WEB&doi=1800000001 (accessed December 2007).
38. Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., Weitzner, D. (2006) Creating a science of the Web. Science 313(5788), 769-771.

39. Berners-Lee, T, Hendler, J., Lassila, O. (2001) The Semantic Web. Scientific American, http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 (accessed December 2007).

40. Boley, H., Kifer, M. (2007) RIF basic logic dialect. http://www.w3.org/TR/rif-bld/ (accessed December 2007).

41. Bonatti, P.A., Duma, C., Fuchs, N., Nejdl, W., Olmedilla, D., Peer, J., Shahmehri, N. (2006) Semantic Web policies – a discussion of requirements and research issues. In Sure, Y., Domingue, J. (Eds.) The Semantic Web: research and applications, 3rd European Semantic Web Conference 2006 (ESWC-06), Budva, Montenegro. Springer, Berlin.

42. Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H. (2004) Evolving GATE to meet new challenges in language engineering. Natural language engineering 10(3/4), 349-373.

43. Bray, T. Paoli, J. Sperberg-McQueen, C.M. Maler, E. Yergeau, F. (2006) Extensible Markup Language (XML) 1.0 (Fourth Edition). http://www.w3.org/TR/xml/, 2006 (accessed December 2007).

44. Brewster, C., Ciravegna, F., Wilks, Y. (2002) User-centred ontology learning for knowledge management. In Andersson, B., Bergholtz, M., Johannesson, P. (Eds.) Proceedings of the 6th international conference on applications of natural language to information systems. Springer, Berlin, 203-207.

45. Brewster, C., O'Hara, K. (2004) Knowledge representation with ontologies: the present and future. IEEE intelligent systems 19(1), 72-81.

46. Brickley, D., Guha, R.V., McBride, B. (2004) RDF vocabulary description language 1.0: RDF Schema. http://www.w3.org/TR/rdf-schema/ (accessed December 2007).

47. Ciravegna, F., Dingli, A., Guthrie, L, Wilks, Y. (2003) Integrating information to bootstrap information extraction from Web sites. In: IJCAI 2003 Workshop on Information Integration on the Web, in conjunction with the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), Acapulco, Mexico.

48. Ciravegna F., Wilks, Y. (2003) Designing adaptive information extraction for the Semantic Web in Amilcare. In Handschuh, S., Staab, S. (Eds.) (2003) Annotation for the Semantic Web. IOS Press, Amsterdam.

49. Codd, E.F. (1970) A relational model of data for large shared data banks. Communications of the ACM 13(6): 377-387.

50. Connolly, D. (2003) Semantic Web update: OWL and beyond. http://www.w3.org/2003/Talks/1017-swup/all.htm (accessed December 2007).

51. Connolly, D. (Ed.) (2007) Gleaning Resource Descriptions from Dialects of Langages (GRDDL). http://www.w3.org/TR/grddl/ (accessed December 2007).

52. De Roure, D., Sure, Y. (Eds.) (2006) The Semantic Grid. Special issue of the Journal of Web Semantics 4(2), http://www.websemanticsjournal.org/navigation.html#4 (accessed December 2007).

53. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V.C., Sachs, J. (2004) Swoogle: a search and metadata engine for the Semantic Web. Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, ACM Press, http://ebiquity.umbc.edu/_file_directory_/papers/115.pdf (accessed December 2007).

54. Ding, L., Zhou, L., Finin, T., Joshi, A. (2005) How the Semantic Web is being used: an analysis of FOAF documents. In Proceedings of the 38th international conference on system sciences. http://ebiquity.umbc.edu/_file_directory_/papers/120.pdf (accessed December 2007).

55. Duerst, M., Suignard, M. (2005) Internationalized Resource Identifiers (IRIs). http://www.ietf.org/rfc/rfc3987.txt (accessed December 2007).

56. Fensel, D. (2004) Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, 2nd Ed. Springer, Berlin.

57. Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F. (2001) OIL: an ontology infrastructure for the Semantic Web. IEEE Intelligent Systems 16(2), 38-45, http://www.cs.man.ac.uk/~horrocks/Publications/download/2001/IEEE-IS01.pdf (accessed December 2007).

58. Fielding, R., Irvine, U.C., Gettys, J., Mogul, J.C., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T. (1999) Hypertext Transfer Protocol – HTTP/1.1. http://www.w3.org/Protocols/HTTP/1.1/rfc2616.pdf (accessed December 2007).

59. Goble, C., Kesselman, C., Sure, Y. (Eds.) (2005) Proceedings of the Dagstuhl seminar on Semantic Grid: the convergence of technologies. http://drops.dagstuhl.de/portals/index.php?semnr=05271 (accessed December 2007).

60. Goguen, J.A. (2004) Ontology, ontotheology and society. In International conference on formal ontology in information systems (FOIS 2004). http://charlotte.ucsd.edu/users/goguen/pps/fois04.pdf (accessed December 2007).

61. Golbeck, J., Hendler, J. (2004) Accuracy of metrics for inferring trust and reputation in Semantic Web-based social networks. In Motta, E., Shadbolt, N., Stutt, A., Gibbins, N. (Eds.) Engineering Knowledge in the Age of the Semantic Web, Proceedings of 14th International Conference, EKAW 2004, Whittlebury Hall, United Kingdom. Springer, Berlin, 116-131.

62. Groth, P., Jiang, S., Miles, S., Munroe, S., Tan, V., Tsasakou, S., Moreau, L. (2006) An architecture for provenance systems. http://eprints.ecs.soton.ac.uk/13216/1/provenanceArchitecture10.pdf (accessed December 2007).

63. Handschuh, S., Staab, S. (Eds.) (2003) Annotation for the Semantic Web. IOS Press, Amsterdam.

64. Handschuh, S., Staab, S., Ciravegna, F. (2002) S-CREAM – Semi-automatic CREAtion of Metadata. In Gómez-Pérez, A., Benjamins, V.R. (Eds.)

Knowledge engineering and knowledge management: ontologies and the Semantic Web, proceedings of 13th international conference, EKAW 2002, Siguënza, Spain. Springer, Berlin, 358-372.

65. Harnad, S. (1990) The symbol grounding problem. Physica D 42, 335-346; http://users.ecs.soton.ac.uk/harnad/Papers/Harnad/harnad90.sgproblem.html (accessed December 2007).
66. Harris, S., Gibbins, N. (2003) 3store: efficient bulk RDF storage. In Proceedings of the 1st International Workshop on Practical and Scalable Systems, Sanibel Island, Florida. http://km.aifb.uni-karlsruhe.de/ws/psss03/proceedings/harris-et-al.pdf (accessed December 2007).
67. Haugeland, J. (1979) Understanding natural language. Journal of philosophy 76, 619-632.
68. Heath, T., Motta, E. (2007) Revyu.com: a reviewing and rating site for the Web of data. In proceedings of the 6th international Semantic Web conference 2007, Busan, South Korea. http://iswc2007.semanticweb.org/papers/889.pdf (accessed December 2007).
69. Heflin, J. (2004) OWL Web Ontology Language use cases and requirements. http://www.w3.org/TR/webont-req/ (accessed December 2007).
70. Hendler, J. (2007) Shirkyng my responsibility. http://www.mindswap.org/blog/2007/11/21/shirkyng-my-responsibility/ (accessed December 2007).
71. Hendler, J., de Roure, D. (2004) E-science: the grid and the Semantic Web. IEEE intelligent systems 19(1), 65-71.
72. Hendler, J., Golbeck, J. (2008) Metcalfe's law applies to Web 2.0 and the Semantic Web. Journal of Web semantics 6(1), http://www.websemanticsjournal.org/papers/2007119/MetcalfsLawGolbeckV6I1.pdf (accessed December 2007).
73. Hu, B., Dasmahapatra, S., Dupplaw, D., Lewis, P., Shadbolt, N. (2007) Reflections on a medical ontology. International journal of human-computer studies 65(7), 569-582.
74. Huang Z., Stuckenschmidt, H. (2005) Reasoning with multi-version ontologies: a temporal logic approach. Proceedings of the 4th International Semantic Web Workshop, http://www.cs.vu.nl/~heiner/public/ISWC05a.pdf (accessed December 2007).
75. Iria, J., Ciravegna F. (2005). Relation extraction for mining the Semantic Web. Dagstuhl seminar on machine learning for the Semantic Web, http://tyne.shef.ac.uk/t-rex/pdocs/dagstuhl.pdf (accessed December 2007).
76. Jones, K.S. (2004) What's new about the Semantic Web? Some questions. SIGIR forum, 38(2), http://www.sigir.org/forum/2004D/sparck_jones_sigirforum_2004d.pdf (accessed December 2007).
77. Kalfoglou, Y., Schorlemmer, M. (2003) Ontology mapping: the state of the art. Knowledge engineering review 18(1), 1-31.

78. Kilgarrif, A., Grefenstette G. (2003). Introduction to the special issue on the Web as corpus. Computational linguistics 29(3), 333-348, http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf (accessed December 2007).

79. Klyne, G., Carroll, J.J. McBride, B. (2004) Resource Description Framework (RDF): concepts and abstract syntax. http://www.w3.org/TR/rdf-concepts/ (accessed December 2007).

80. Lenat, D.B. (1995) Cyc: a large-scale investment in knowledge infrastructure. Communications of the ACM 38(11).

81. Leonard, T., Glaser, H. (2001) Large scale acquisition and maintenance from the Web without source access. In proceedings of workshop on knowledge markup and semantic annotation, K-CAP2001. http://eprints.ecs.soton.ac.uk/6185/1/Paper.pdf (accessed December 2007).

82. Manola, F., Miller, E., McBride, B. (2004) RDF primer. http://www.w3.org/TR/rdf-primer/ (accessed December 2007).

83. McGuinness, D.L., Fikes, R., Stein, L.A., Hendler, J. (2003) DAML-ONT: an ontology language for the Semantic Web. In: Fensel, D., Hendler, J., Lieberman, H., Wahlster, W. (eds.) Spinning the Semantic Web: bringing the World Wide Web to its full potential. MIT Press, Cambridge, MA, 65-93.

84. McGuinness, D.L., van Harmelen, F. (2004) OWL Web Ontology Language overview. http://www.w3.org/TR/owl-features/ (accessed December 2007).

85. Mika, P. (2005) Flink: Semantic Web technology for the extraction and analysis of social networks. Journal of Web semantics 3(2). http://www.websemanticsjournal.org/papers/20050719/document7.pdf (accessed December 2007).

86. Mika, P. (2007) Ontologies are us: a unified model of social networks and semantics. Journal of Web semantics 5(1), 5-15.

87. Nowack, B. (2005) CONFOTO: A semantic browsing and annotation service for conference photos. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (Eds.) The Semantic Web, proceedings of the international Semantic Web conference 2005, Hiroshima, Japan. Springer, Berlin, 1067-1070.

88. Noy, N.F., McGuinness, D.L. (2001) Ontology development 101: a guide to creating your first ontology. http://smi.stanford.edu/smi-web/reports/SMI-2001-0880.pdf (accessed December 2007).

89. Noy, N.F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R.W., Musen, M.A. (2001). Creating Semantic Web contents with Protégé-2000. IEEE Intelligent Systems 16(2), 60-71.

90. O'Hara, K. (2004) Ontologies and technologies: knowledge representation or misrepresentation. SIGIR forum, 38(2), http://sigir.org/forum/2004D/ohara_sigirforum_2004d.pdf (accessed December 2007).

91. O'Hara, K., Alani, H., Kalfoglou, Y., Shadbolt, N. (2004) Trust strategies for the Semantic Web. In Workshop on trust, security and reputation on the Semantic Web, 3rd international Semantic Web conference (ISWC 04),

Hiroshima, Japan. http://eprints.ecs.soton.ac.uk/10029/ (accessed December 2007).

92. O'Hara, K., Shadbolt, N. (2008) The spy in the coffee machine: the end of privacy as we know it. Oneworld, Oxford.

93. Patel-Schneider, P., Horrocks, I., van Harmelen, F. (2002) Reviewing the design of DAML+OIL: an ontology language for the Semantic Web. In Proceedings of the 18th National Conference on Artificial Intelligence (AAAI02), Edmonton, Canada. http://www.cs.vu.nl/~frankh/postscript/AAAI02.pdf (accessed December 2007).

94. Pike, W., Gahegan, M. (2007) Beyond ontologies: toward situated representations of scientific knowledge. International journal of human-computer studies 65(7), 674-688.

95. Prud'hommeaux, E., Seaborne, A. (2007) SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query/ (accessed December 2007).

96. Putnam, H. (1975) The meaning of 'meaning'. Philosophical papers vol.2: mind, language and reality. Cambridge University Press, Cambridge.

97. schraefel, m.m.c., Shadbolt, N.R., Gibbins, N., Glaser, H., Harris, S. (2004) CS AKTive Space: representing computer science on the Semantic Web. In Proceedings of WWW 2004, New York, 2004. http://eprints.ecs.soton.ac.uk/9084/ (accessed December 2007).

98. Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., de Niet, M., Omelayenko, B., van Ossenbruggen, J., Siebes, R., Taekema, J., Wielemaker, J., Wielinga, B. (2006) MultimediaN e-culture demonstrator. http://www.cs.vu.nl/~guus/papers/Schreiber06a.pdf (accessed December 2007).

99. Schröder, M., Zovato, E., Pirker, H., Peter, C., Burkhardt, F. (2007) W3C emotion incubator group report. http://www.w3.org/2005/Incubator/emotion/XGR-emotion/ (accessed December 2007).

100. Shadbolt, N., Hall, W., Berners-Lee, T. (2006) The Semantic Web revisited. IEEE Intelligent Systems, *21* (3), 96-101.

101. Shafer, G. (1998). 'Causal logic', Proceedings of IJCAI-98, http://www.glennshafer.com/assets/downloads/articles/article62.pdf (accessed December 2007).

102. Shirky, C. (2005) Ontology is overrated: categories, links and tags. http://www.shirky.com/writings/ontology_overrated.html (accessed December 2007).

103. Stevens, R., Egaña Aranguren, M., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., Rector, A. (2007) Using OWL to model biological knowledge. International journal of human-computer studies 65(7), 583-594.

104. Stojanovic, L., Stojanovic, N., Volz, R. (2002) Migrating data-intensive Web sites into the Semantic Web. Symposium on Applied Computing, Madrid.

105. Troncy, R., van Ossenbruggen, J., Pan, J.Z., Stamou, G., Halaschek-Wiener, C., Simou, N., Tsouvaras, V. (2007) Image annotation on the Semantic Web. http://www.w3.org/2005/Incubator/mmsem/XGR-image-annotation/ (accessed December 2007).

106. Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., Ciravegna, F. (2002) MnM: ontology-driven semi-automatic and automatic support for semantic markup. In Gómez-Pérez, A., Benjamins, V.R. (Eds.) Knowledge engineering and knowledge management: ontologies and the Semantic Web, proceedings of 13th international conference, EKAW 2002, Siguënza, Spain. Springer, Berlin, 379-391.

107. Weitzner, D.J., Hendler, J., Berners-Lee, T., Connolly, D. (2005) Creating a policy-aware Web: discretionary, rule-based access for the World Wide Web. In Ferrari, E., Thuraisingham, B. (Eds.) Web and information security. Idea Group Inc, Hershey, PA.

108. Wilks, Y. (2006) The Semantic Web as the apotheosis of annotation, but what are its semantics? http://www.dcs.shef.ac.uk/~lucy/yw_pubs/yorick-wilks-semantic-annotation.pdf (accessed December 2007).

109. Wittgenstein, L. (1953) Philosophical investigations. Basil Blackwell, Oxford.

110. Zambonini, D. (2006) The 7 (f)laws of the Semantic Web. http://www.oreillynet.com/xml/blog/2006/06/the_7_flaws_of_the_semantic_we.html *[sic]* (accessed December 2007).

Books and reviews

Asdasdfadsasdasd

1. Antoniou, G., van Harmelen, F. (2004) A Semantic Web primer. MIT Press, Cambridge, MA.

2. Berners-Lee, T. (1999) Weaving the Web: the past, present and future of the World Wide Web by its inventor. Texere Publishing, London.

3. Berners-Lee, T., Hall, W., Hendler, J.A., O'Hara, K., Shadbolt, N., Weitzner, D.J. (2006) A framework for Web Science. Foundations and Trends in Web Science 1(1), 1-134, http://www.nowpublishers.com/product.aspx?product=WEB&doi=1800000001 (accessed December 2007).

4. Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., Weitzner, D. (2006) Creating a science of the Web. Science 313(5788), 769-771.

5. Berners-Lee, T, Hendler, J., Lassila, O. (2001) The Semantic Web. Scientific American, http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 (accessed December 2007).

6. Fensel, D. (2004) Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, 2nd Ed. Springer, Berlin.

7. Fensel, D., Hendler, J., Lieberman, H., Wahlster, W. (Eds.) (2003) Spinning the Semantic Web: bringing the World Wide Web to its full potential. MIT Press, Cambridge, MA.
8. Shadbolt, N., Hall, W., Berners-Lee, T. (2006) The Semantic Web revisited. IEEE Intelligent Systems, *21* (3), 96-101.