Foundations and Trends® in
Web Science
Vol. 1, No 1 (2006) 1–130

DOI: 10.1561/1800000001

now
the essence of knowledge

A Framework for Web Science

Tim Berners-Lee[1], Wendy Hall[2], James A. Hendler[3], Kieron O'Hara[4], Nigel Shadbolt[4] and Daniel J. Weitzner[5]

[1] *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology*
[2] *School of Electronics and Computer Science, University of Southampton*
[3] *Department of Computer Science, Rensselaer Polytechnic Institute*
[4] *School of Electronics and Computer Science, University of Southampton*
[5] *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology*

Abstract

This text sets out a series of approaches to the analysis and synthesis of the World Wide Web, and other web-like information structures. A comprehensive set of research questions is outlined, together with a sub-disciplinary breakdown, emphasising the multi-faceted nature of the Web, and the multi-disciplinary nature of its study and development. These questions and approaches together set out an agenda for *Web Science*, the science of decentralised information systems. Web Science is required both as a way to understand the Web, and as a way to focus its development on key communicational and representational requirements. The text surveys central engineering issues, such as the development of the Semantic Web, Web services and P2P. Analytic approaches to discover the Web's topology, or its graph-like structures, are examined. Finally, the Web as a technology is essentially socially embedded; therefore various issues and requirements for Web use and governance are also reviewed.

1

Introduction

The World Wide Web is a technology that is only a few years old, yet its growth, and its effect on the society within which it is embedded, have been astonishing. Its inception was in support of the information requirements of research into high energy physics. It has spread inexorably into other scientific disciplines, academe in general, commerce, entertainment, politics and almost anywhere where communication serves a purpose [142, 143]. Freed from the constraints of printing and physical distribution, the results of scientific research, and the data upon which that research is carried out, can be shared quickly. Linking allows the work to be situated within rich contexts. Meanwhile, innovation has widened the possibilities for communication. Weblogs and wikis allow the immediacy of conversation, while the potential of multimedia and interactivity is vast.

But neither the Web nor the world is static. The Web evolves in response to various pressures from science, commerce, the public and politics. For instance, the growth of e-science has created a need to integrate large quantities of diverse and heterogeneous data; e-government and e-commerce also demand more effective use of information [34]. We need to understand these evolutionary and developmental forces.

Without such an appreciation opportunities for adding value to the Web by facilitating more communicative and representational possibilities may be missed. But development is not the whole of the story. Though multi-faceted and extensible, the Web is based on a set of architectural principles which need to be respected. Furthermore, the Web is a social technology that thrives on growth and therefore needs to be trusted by an expanding user base – trustworthiness, personal control over information, and respect for the rights and preferences of others are all important aspects of the Web. These aspects also need to be understood and maintained as the Web changes.

A research agenda that can help identify what needs to stay fixed and where change can be profitable is imperative. This is the aim of *Web Science*, which aims to map how decentralised information structures can serve these scientific, representational and communicational requirements, and to produce designs and design principles governing such structures [34]. We contend that this science of decentralised information structures is essential for understanding how informal and unplanned informational links between people, agents, databases, organisations and other actors and resources can meet the informational needs of important drivers such as e-science and e-government. How an essentially decentralised system can have such performance designed into it is the key question of Web Science [34].

'Web Science' is a deliberately ambiguous phrase. Physical science is an analytic discipline that aims to find laws that generate or explain observed phenomena; computer science is predominantly (though not exclusively) synthetic, in that formalisms and algorithms are created in order to support particular desired behaviour. Web science has to be a merging of these two paradigms; the Web needs to be *studied* and understood, and it needs to be *engineered*. At the micro scale, the Web is an infrastructure of artificial languages and protocols; it is a piece of engineering. But the linking philosophy that governs the Web, and its use in communication, result in emergent properties at the macro scale (some of which are desirable, and therefore to be engineered in, others undesirable, and if possible to be engineered out). And of course the Web's use in communication is part of a wider system of human interaction governed by conventions and laws. The various levels at which Web

technology interacts with human society mean that interdisciplinarity is a firm requirement of Web Science.

Such an interdisciplinary research agenda, able to drive Web development in socially and scientifically useful ways, is not yet visible and needs to be created. To that end, in September 2005 a Web Science Workshop was convened in London, UK (details of the contributors to the Workshop are given in the Acknowledgements). The workshop examined a number of issues, including:

- Emerging trends on the Web.
- Challenges to understanding and guiding the development of the Web.
- Structuring research to support the exploitation of opportunities created by (*inter alia*) ubiquity, mobility, new media and the increasing amount of data available online.
- Ensuring important social properties such as privacy are respected.
- Identifying and preserving the essential invariants of the Web experience.

This text grew out of the Web Science Workshop, and it attempts to summarise, expand and comment on the debates. That an interdisciplinary approach was required was agreed by all, encompassing computer science and engineering, the physical and mathematical sciences, social science and policymaking. Web Science, therefore, is not just about methods for modelling, analysing and understanding the Web at the various micro- and macroscopic levels. It is also about engineering protocols and providing infrastructure, and ensuring that there is fit between the infrastructure and the society that hosts it. Web Science must coordinate engineering with a social agenda, policy with technical constraints and possibilities, analysis with synthesis – it is inherently interdisciplinary, and this text is structured to reflect that.

Developing the Web also involves determining what factors influence the Web experience, and ensuring that they remain in place. Examples of basic architectural decisions that underpin the Web include: the 404 error, which means that failure to link to a resource doesn't cause catastrophic failure; the use of the Uniform Resource Indicator (URI); and

the full exploitation of the pre-existing Internet infrastructure (such as the Domain Name System) as the platform on which the Web was built. Standards are also crucial, and the World Wide Web Consortium's (W3C) work of creating and recommending standards while maintaining stakeholder consensus shows that engineering needs to go hand in hand with a social process of negotiation.

Section 2 reviews these basic scientific and architectural principles in more detail. Exploring the metaphor of 'evolution' may help us to envisage the Web as a populated ecology, and as a society with the usual social requirements of policies and rules. Connecting relevant approaches, covering variant methodologies, varying spatiotemporal grain sizes and modelling across a wide range of domains, will be challenging.

Section 3 looks at some of the issues to do with engineering the Web, and how to promote, and be promoted by, new technologies such as grids or services. Perhaps one of the most important potential developments to be discussed in this section is the Semantic Web. The Web is usually characterised as a network of linked documents many of which are designed to be read by humans, so that machine-readability requires the heuristics of natural language processing. However, the Semantic Web, a vision of extending and adding value to the Web, is intended to exploit the possibilities of logical assertion over linked relational data to allow the automation of much information processing. Research and development has been underway for some time now on developing the languages and formalisms that will support querying, inference, aligning data models, visualisation and modelling.

To flourish, the Semantic Web needs the same decentralising philosophy as the World Wide Web. One challenge is to ensure that various individual data systems can be amalgamated with local consistency without attempting the impossible task of trying to enforce consistency globally. Furthermore, the basic use of a common set of symbols – URIs – by a number of formalisms with contrasting properties, such as rules and logic, without assuming any kind of centralised or 'basic' formalism for describing the Web is also non-trivial. A third issue is to do with bringing data together to leverage the power of amalgamation and serendipitous reuse; most data currently sit in standalone repositories

and are not published (in contrast to the WWW, where documents are routinely made available to a wider audience).

Section 4 looks at attempts to analyse the Web in ways that can feed back into the engineering effort. For instance, modelling the Web mathematically will enable search and information retrieval to keep pace with its growth, especially if linked to important fields such as natural language processing, network analysis and process modelling. Understanding emergent structures and macroscopic topology will help to generate the laws of connectivity and scaling to which the Web conforms.

As noted, the Web's value depends on its use by and in society, and its ability to serve communication needs without destroying other valuable types of interaction. This means understanding those needs, their relation to other social structures, and the two-way interaction with technological development. Social issues such as these are discussed in Section 5, and include philosophical issues to do with the meaning of symbols, logical problems such as methods of reasoning, and social issues including the creation and maintenance of trust, and the mapping of social communities via their activities on the Web.

Some of the interactions between society and Web technology are current and require policies for regulation and expressing preferences. For instance, the Semantic Web clearly motivates a corporate and individual cultural imperative to publish and share data resources, which in turn will require policies dealing with access control, privacy, identity and intellectual property (as well as interfaces and systems that can express policy rules to a heterogeneous user base). Policy, governance and political issues such as these are discussed in Section 6.

Section 7 provides a brief conclusion, summarising the case for a Science of the Web, and encapsulating the vision that this text, in an extended form, has presented.

2

The Web and its Science

We may paraphrase Web Science as the science of the Web. Whilst this equivalence may be obvious we will begin by breaking down the phrase and sketching the components that enable the Web to function as an effective decentralised information system. We will review the basic architectural principles of the Web, designed to support growth and the social values of information-sharing and trustworthy behaviour in Section 2.1. Section 2.2 will then offer a few methodological reflections on the scientific investigation of the Web.

2.1 Web architecture

The architecture of the Web exploits simple technologies which connect efficiently, to enable an information space that is highly flexible and usable, and which, most importantly, scales. The Web is already an impressive platform upon which thousands of flowers have bloomed, and the hope is it can grow further, encompassing more languages, more media and more activities, hosting more information, as well as providing the tools and methods for interrogating the data that is out there. In this opening section we will briefly review the main principles

underlying Web architecture; this section is indebted to [155], and for more detail see that document.

The Web is a space in which *resources* are identified by *Uniform Resource Identifiers* (URIs – [33]). There are *protocols* to support *interaction* between agents, and *formats* to *represent* the information resources. These are the basic ingredients of the Web. On their design depends the utility and efficiency of Web interaction, and that design depends in turn on a number of principles, some of which were part of the original conception, while others had to be learned from experience.

Identification of resources is essential in order to be able to share information about them, reason about them, modify or exchange them. Such resources may be anything that can be linked to or spoken of; many resources are purely information, but others not. Furthermore, not all resources are on the Web, in that they may be *identifiable* from the Web, but may not be *retrievable* from it. Those resources which are essentially information, and which can therefore be rendered without abstraction and characterised completely in a message are called *information resources*.

For these reasoning and referring functions to happen on the global scale, an identification system is required to provide a single global standard; URIs provide that system. It would be possible for alternative systems to URIs to be developed, but the added value of a *single* global system of identifiers, allowing linking, bookmarking and other functions across heterogeneous applications, is high. Resources have URIs associated with them, and each URI ideally identifies a single resource in a context-independent manner. URIs act as names (and addresses – see Section 3.1.2 below for discussion of this issue), and so if it is possible to guess the nature of a resource from its URI, that is a contingent matter; in general URIs refer opaquely. These principles of relationship between URIs and resources are desirable but not strictly enforceable; the cost of failing to associate a URI with a resource is the inability to refer to it, while the cost of assigning two resources to a URI will be error, as data about one resource gets applied to the other.

URIs also connect the Web with the offline social world, in that they require institutions. They fall under particular defined schemes,

of which perhaps the most commonly understood are HTTP, FTP and mailto; such schemes are registered with the Internet Assigned Numbers Authority (IANA – http://www.iana.org/assignments/uri-schemes). These schemes need to be operated on principled lines in order to be effective.

So if we take HTTP as an example, HTTP URIs are owned and disbursed by people or organisations; and hence can be allocated responsibly or irresponsibly. For instance, an HTTP URI should refer to a single resource, and be allocated to a single owner. It is also desirable for such a URI to refer to its resource permanently, and not change its reference over time (see Section 5.4.6 below). Communication over the Web involves the exchange of messages which may contain data or metadata about a resource. One common aim of communication is to access a resource via a URI, or to *dereference* the URI. If a resource has been given an identifier, the resource has to be in some way recoverable from the identifier for it to be of value. Dereferencing typically involves finding an appropriate index to look up the identifier. There are often clues in the identifier, or the use of the identifier, that help here, particularly if the naming authorities have some kind of hierarchical structure.

For example, a postal address has a hierarchical structure that enables a particular building to be located. One would consult the index of the London A-Z to find a particular street whose name one knew and which one knew was located in London but nothing further about it. Similarly, the Domain Name System (DNS) exploits hierarchical structure to help with dereferencing, so that to contact the server "foo.cs.bar.edu" involves sending a message of some sort to some server controlled by Bar University in the United States. The more information that is available in the name or identifier of a resource, the easier it is to dereference, although of course in the limiting case a resource's name need contain no information at all to aid in dereferencing it (sometimes this is the point of a name). Furthermore, identification systems often need to be maintained by authorities for dereferencing to be possible – if the London A-Z was not updated every so often, it would become impossible to use it (the latest edition) to find particular houses, in the same way that changes in Bar University's server

maintenance programme could mean that some resources held on its servers were unlocatable.

What accessing an information resource entails varies from context to context, but perhaps the most common experience is receiving a representation of the (state of the) resource on a browser. It certainly need not be the case that dereferencing a URI automatically leads to the agent getting privileged access to a resource. It may be that no representation of the resource is available, or that access to a resource is secure (e.g. password controlled), but it should be possible to refer to a resource using its URI without exposing that resource to public view. The development of the Web as a space, rather than a large and complex notice board, follows from the ability of agents to use interactions to alter the states of resources, and to incur obligations and responsibilities. Retrieving a representation is an example of a so-called *safe* interaction where no alteration occurs, while posting to a list is an *unsafe* interaction where resources' states may be altered. Note that the universal nature of URIs helps the identification and tracking of obligations incurred online through unsafe interactions.

Not all URIs are intended to provide access to representations of the resources they identify. For instance, the mailto: scheme identifies resources that are reached using Internet mail (e.g. mailto:romeo@example.edu identifies a particular mailbox), but those resources aren't *recoverable* from the URI in the same way as a webpage is. Rather, the URI is used to *direct* mail to that particular mailbox, or alternatively to find mail from it.

The Web supports a wide variety of file formats, of which the most well-known is HTML. Several formats are required, and formats need to be flexible, because of the heterogeneous nature of interaction over the Web. Content may be accessed via all sorts of devices, most commonly a PC or a mobile device, and more value can be extracted from the Web if the presentation of content is device-independent as far as possible (ideally compatible with devices not yet dreamt of). Separating the representation of content from the concerns of presentation and interaction is good practice here; under such a regime, content, presentation and interaction need to be recombined in a way that is maximally

useful, which is generally done in part by the server and in part by the client, the exact ratio between the two depending on the context of the interaction.

The power of the Web stems from the linking it makes possible. A resource can contain a reference to another resource in the form of an embedded URI which can be used to access the second resource. These links allow associative navigation of the Web. To facilitate linking, a format should include ways to create and identify links to other resources, should allow links to any resources anywhere over the Web, and should not constrain content authors to using particular URI schemes.

An important aim of Web Science is to identify the essential aspects of identification, interaction and representation that make the Web work, and to allow the implementation of systems that can support or promote desirable behaviour. The experience of linking documents and, increasingly, data releases great power, both for authors and users. The possibility of serendipitous reuse of content empowers authors by increasing their influence, and users by providing access to much more information than would be possible using other technologies.

In particular, the three functions of identification, interaction and representation need to be separated out. Altering or adding a scheme for identification, say, should have no effect on schemes for interaction or representation, allowing independent, modular evolution of the Web architecture as new technologies and new applications come on stream (which is not to say that orthogonal specifications might not co-evolve cyclically with each other). Similarly, technologies should be extensible, that is they should be able to evolve separately without threatening their interoperability with other technologies.

Finally, it is an essential principle of Web architecture that errors should be handled simply and flexibly. Errors are essential – in an information space whose size can be measured in thousands of terabytes, and the numbers of users in the hundreds of millions, heterogeneity of purpose and varying quality of authorship mean that there will be design errors aplenty. The existence of dangling links (links using a URI with no resource at the end of it), non-well-formed content or other predictable errors should not cause the system to crash; the demands

of interoperability require that agents should be able to recover from errors, without, of course, compromising user awareness that the error has occurred.

As the Web grows and develops to meet new situations and purposes, the architecture will have to evolve. But the evolution needs to be gradual and careful (the slow and necessarily painstaking negotiations of standards committees are a good way to combine gradualism with fitness for purpose), and the principle of keeping orthogonal developments separate means that evolution in one area should not affect evolution elsewhere. The evolution needs to respect the important invariants of the Web, such as the URI space, and it is important that developers at all times work to preserve those aspects of the Web that need to be preserved. This is part of the mission of the W3C's Technical Architecture Group [154], although standards can only ever be part of the story. Web architectural principles will always be debated outside the W3C, quite properly, as well as within it.

2.2 Web science: Methodology

If the investigation of the Web is to be counted as properly scientific, then an immediate question is how scientific method should apply to this particular domain. How should investigators and engineers approach the Web in order to understand it and its relation to wider society, and to innovate?

Various aspects of the Web are relatively well-understood, and as an engineered artefact its building blocks are crafted, not natural phenomena. Nevertheless, as the Web has grown in complexity and the number and types of interactions that take place have ballooned, it remains the case that we know more about some complex natural phenomena (the obvious example is the human genome) than we do about this particular engineered one.

However it actually evolves, any Web Science deserving of the name would need to meet some obvious conditions. There would need to be falsifiabilty of hypotheses and repeatability of investigations. There would need to be independent principles and standards for assessing when a hypothesis had been established. There is a real issue as to

how these principles and standards should be arrived at. And of course there should be methods for moving from assessments of the Web and its evolution to the development and implementation of innovation.

To take one example, there are a number of technologies and methods for mapping the Web and marking out its topology (see Section 4.1 below). What do such maps tell us (cf. e.g. [80])? The visualisations are often very impressive, with three-dimensional interpretations and colour-coded links between nodes. But how verifiable are such maps? In what senses do they tell us 'how the Web is'? What are the limitations?

The obvious application, in methodological terms, of maps and graphs of the Web's structure is to direct sampling, by specifying the properties that models and samples of the Web should have. The rapid growth of the Web made a complete survey out of the question years ago, and information scientists need rapid and timely statistics about the content of Web-available literature. Representative sampling is key to such methods, but how should a sample be gathered in order to be properly called representative [188]? To be properly useful, a sample should be *random*; 'randomness' is usually defined for particular domains, and in general means that all individuals in the domain have an equal probability of being selected for the sample. But for the Web that entails, for example, understanding what the individuals are; for instance, are we concerned with websites or webpages? If the former, then one can imagine difficulties as there is no complete enumeration of them. And sampling methods based on, say, IP addresses are complicated by the necessarily sparse population of the address space [219].

Furthermore, so cheap are operations on the Web that a small number of operators could skew results however carefully the sample is chosen. A survey reported in more detail below [99] apparently discovered that 27% of pages in the .de domain changed every week, as compared with 3% for the Web as a whole. The explanation turned out not to be the peculiar industriousness of the Germans, but rather over a million URLs, most but not all on German servers, which resolved to a single IP address, an automatically-generated and constantly-changing pornography site.

The Web has lots of unusual properties that make sampling trickier; how can a sampling method respect what seem *prima facia* significant

properties such as, for instance, the percentage of pages updated daily, weekly, etc? How can we factor in such issues as the independence of underlying data sources? Do we have much of a grasp of the distribution of languages across the Web (and of terms within languages – cf. [167]), and how does the increasing cleverness in rendering affect things [138]? And even if we were happy with our sampling methodology, how amidst all the noise could we discover interesting structures efficiently [191]?

Furthermore, although for many purposes the Web can be treated as a static information space, it is of course dynamic and evolving. So any attempt at longitudinal understanding of the Web would need to take that evolution into account [218], and models should ideally have the growth of the system (in terms of constant addition of new vertices and edges into the graph), together with a link structure that is not invariant over time, and hierarchical domain relationships that are constantly prone to revision, built into them (cf. e.g. [253]).

Analytic modelling combined with carefully collected empirical data can be used to determine the probabilities of webpages being edited (altering their informational content) or being deleted. One experiment of observation of hundreds of thousands of pages over several months produced interesting results: at any one time round about 20% of webpages were under 11 days old, while 50% appeared in the previous three months. On the other hand, 25% were over a year old – age being defined here as the difference between the time of the last modification to the page and the time of downloading [43]. Another experiment involved crawling over 150m HTML pages once a week for 11 weeks, and discovered, for example, strong connections between the top-level domain and the frequency of change (.com pages changed more frequently than .gov or .edu pages), and that large documents (perhaps counterintuitively) changed more frequently than small ones.

The frequency of past changes was a good predictor of future changes, a potentially important result for incremental Web crawlers [99]. The development of methods of sampling the Web feeds very quickly into the development of more efficient and accurate search. Methods for finding information online, whether logical or heuristic, whether data-centred or on the information retrieval model, require accurate mapping.

So one aspect of Web Science is the investigation of the Web in order to spot threats, opportunities and invariants for its development. Another is the engineering of new, possibly unexpected methods of dealing with information, which create non-conservative extensions of the Web. Such engineering may be research-based, or industry-based. The synthesis of new systems, languages, algorithms and tools is key to the coherent development of the Web, as, for example, with the study of cognitive systems, where much of the progress of the last few years has come with exploratory engineering as well as analysis and description (cf. e.g. [51]). So, for instance, the only way to discover the effects of radically decentralised file sharing is to develop peer to peer systems and observe their operation on increasingly large scales. Such pioneering engineering efforts are vital for the Web's development; it is after all a construction. It is essential for the Web as a whole that implementations of systems interact and don't interfere, which is where standards bodies play an important role.

Hence Web Science is a combination of synthesis, analysis and governance. In the rest of this text, we will take these three aspects in turn, beginning with synthesis, then analysis, and then the social issues that impact on Web development, before finishing off with a discussion of governance issues.

3

Engineering the Web

Tracking the Web's development, determining which innovations are good (e.g. P2P) and which bad (e.g. phishing), and contributing to the beneficial developments are key aims of Web Science. In this section, we will review some of the current directions for builders of the Web. We will look at the Semantic Web and some of the issues and controversies surrounding that (Section 3.1), issues to do with reference and identity (that are important for the Semantic Web to be sure, but also for any type of fruitful information analysis – Section 3.2), and then a selection of further initiatives, including Web services, P2P, grid computing and so on (Section 3.3).

3.1 Web semantics

The Web is a principled architecture of standards, languages and formalisms that provides a platform for many heterogeneous applications. The result might easily be a tangle, and decisions made about the standards governing one formalism can have ramifications beyond, which can of course lead to complex design decisions (cf. [146]). Indeed, the multiple demands on the Web create a temptation to model it

semantically with highly expressive formalisms, but such expressivity generally trades off against usability and a small set of well-understood principles.

However, it is often the case that the trade-off between expressivity and usability is a result of common *mis*use of such formalisms. For instance – we will discuss this example in more detail below – the use of the machinery, implemented and proposed, of the Semantic Web [35, 17] for extending the Web is a common aim. But the design of the SW and its associated formalisms and tools is intended to extend the Web to cover linked data, not, as is often assumed, to improve search or get greater power from annotated text (which is another, separate, type of extension of the Web).

It may be, as many claim and hope, that local models and emergent semantics form an important part of our methods of comprehending the Web. If this is so, there will be a serious trade-off with interoperability: the benefits of structured distributed search and data sharing are large but require interoperable semantics. Leaving semantics underdetermined means forcing the (human) user to do the sense making, as for example with current P2P systems which, if they impose semantics at all, tend only to use very simple, low-level, task-relative structures. In particular, the assumption that the apparatus of the Semantic Web is designed to extend the technologies available for looking at documents can lead to a worry about the trade-off between "easy" emergent semantics and "difficult" logic that is misplaced; we must be careful not to confuse two separate application areas.

3.1.1 The Semantic Web

The Web started life as an attempt to get people to change their behaviour in an important way. Many people create documents, but pre-Web the assumption was that a document was the private property of its creator, and a decision to publish was his or hers alone. Furthermore, the technology to allow people to publish and disseminate documents cheaply and easily was lacking. The Web's aim was to alter that behaviour radically and provide the technology to do it: people would make their documents available to others by adding links

to make them accessible by link following. The rapid growth of the Web, and the way in which this change was quickly adopted in all sectors of Western society have perhaps obscured the radicalism of this step.

The Semantic Web (SW) is an attempt to extend the potency of the Web with an analogous extension of people's behaviour. The SW tries to get people to make their *data* available to others, and to add links to make them accessible by link following. So the vision of the SW is as an extension of Web principles from documents to data. This extension, if it happens and is accepted, will fulfil more of the Web's potential, in that it will allow data to be shared effectively by wider communities, and to be processed automatically by tools as well as manually [34]. This of course creates a big requirement: such tools must be able to process together data in heterogeneous formats, gathered using different principles for a variety of primary tasks. The Web's power will be that much greater if data can be defined and linked so that machines can go beyond display, and instead integrate and reason about data across applications (and across organisational or community boundaries). Currently, the Web does very well on text, music and images, and passably on video and services, but data cannot easily be used on the Web scale [135]. The aim of the SW is to facilitate the *use* of data as well as their discovery, to go beyond Google in this respect.

In this context it is worth mentioning the distinction between *information retrieval* and *data retrieval* (alias automated question-answering). The goal of the former is to produce documents that are relevant to a query; these documents need not be unique, and two successful episodes of information retrieval may nevertheless produce entirely different outcomes. The aim of the latter is to produce the correct answer to the query. There are vast differences between these two types of retrieval, and the stricter adherence to formal principles that the latter project requires may well be a key determinant of what structures one should select when one is finding schemes to provide significance for the terms in one's query. Data are in a very real sense more fundamental than a document; hence the potential increase in the Web's power. There are also a lot of data out there.

A second open issue is exactly what functionality can be achieved by bringing out the relationships between various data sources.

Traditionally, in AI for example, knowledge bases or expert systems, or even databases within an organisation, are used to represent certified information that is reliable, trustworthy, probably consistent and often based on centralised acquisition strategies and representational protocols. On the Web, of course, these assumptions don't necessarily apply. For instance, we must make sure that inconsistencies (which we must expect to encounter on the Web) don't derail all inferences from a particular group of mutually inconsistent knowledge sources. Many of the applications for the SW are yet to come on stream, but some way of coming to terms with the potential scruffiness even of well-structured data from several sources is an issue [278].

The strategy the SW follows, therefore, is to provide a common framework for this liberation of data, based on the Resource Description Framework (RDF), which integrates a variety of applications using XML as the interchange syntax [195]. Raw data in databases are brought together, and connected to models of the world (via ontologies – see below), which then allows the aggregation and analysis of data by producing consistent interpretations across heterogeneous data sources. The focus, therefore, is on the data itself. The SW is not simply a matter of marking up HTML documents on the Web, nor a variant on the traditional IR problem of document retrieval. It is an attempt to bring together data across the Web so as to create a vast database transcending its components, which makes possible applications that infer across heterogeneous data, such as CS AKTive Space which allows browsing and inference across various data sources that chronicle the state of the computer science discipline in the United Kingdom [251].

The SW data model is closely connected with the world of relational data (in which data are represented as n-ary relations, which correspond to a table – [62]), so closely indeed that there is a straightforward mapping from relational databases to RDF. A relational database is a table which is made up of records, which are the rows. Each record is made up of fields, fields being analogous to columns, and an individual record is no more than the contents of its fields (the contents of the cells of the matrix that fall within the rows). Records are RDF nodes, fields are RDF properties and the record field is a value [28].

So, for example, such a table might represent data about cars. Each row (record) would be associated with a particular car, and each column some property or field (colour, owner, registration number, type, recent mechanical history and so on). So some particular property of the car represented in a record would be represented in the appropriate record field. The table might also contain extra information that is harder to express in RDF or in the relational model itself. For instance, Massachusetts State might own a relational database of cars that includes a field for the Massachusetts plate. In that event, the database might be intended to be definitive, i.e. a car is represented in the database if and only if it has a legal Massachusetts plate. That is of course an important property of the table [28].

This sort of database is the type of knowledge source whose exploitation is conceived as the basis for the SW. So the SW is an extension of the WWW in terms of its being the next stage of linking – linking data not documents. It is *not* a set of methods to deal specifically with the documents that are currently on the Web, not a set of inference methods based on metadata or a way of classifying current webpages, or a super-smart way of searching. It is intended to function in the context of the relational model of data.

Linking is key to the SW. In particular, although the publishing of data and the use of RDF is essential, in many cases the practice has been the conversion of data into RDF and its publication divorced from real-world dataflow and management. The languages, methods and tools are still being rolled out for the SW, layer by layer, and it is perhaps unsurprising that quick wins don't appear from the publication of RDF before the tools for viewing, querying and manipulating databases have reached the market. Indeed, as data publication often removes data from its organisational context, the new situation for many will seem worse than the pre-SW era: the application- and organisation-specific tools for manipulating data that had evolved with organisations will have provided a great deal of functionality that may have been lost or eroded. Meanwhile, the lack of linking between data undermines the even greater potential of the SW.

The next layer of the SW is the Web Ontology Language OWL [198], which provides the expressive means to connect data to the world

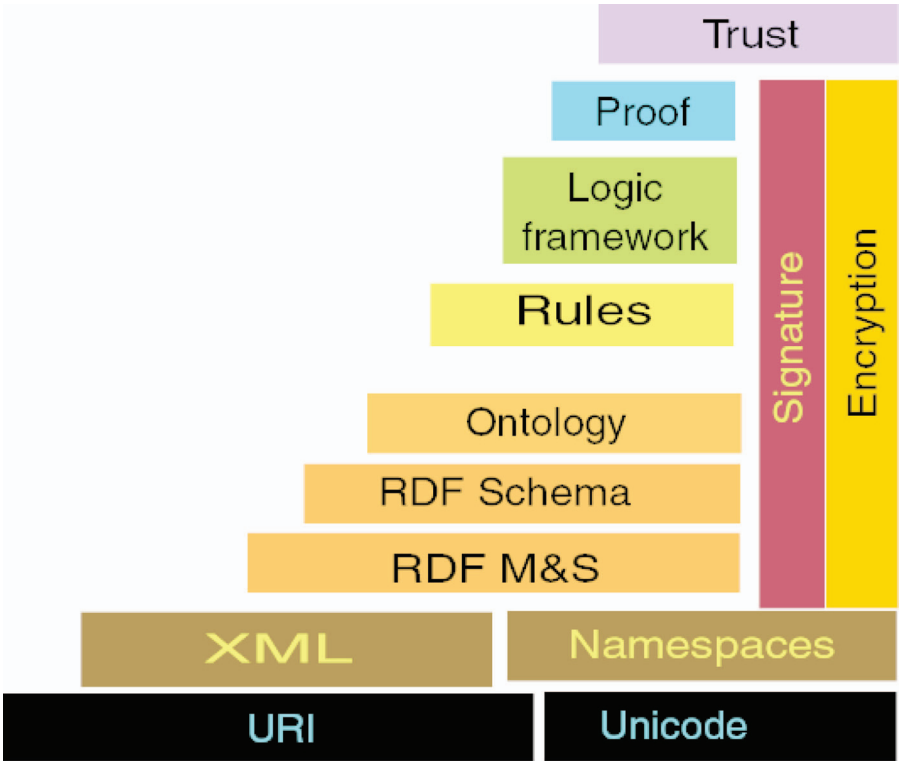

Fig. 3.1 The layers of the Semantic Web.

(as also does RDF Schema or RDF-S – [44]). RDF and OWL allow the exchange of data in a real-world context; on top of this core will sit a query language for RDF which will allow distributed datasets to be queried in a standardised way and with multiple implementations. SPARQL enables the interrogation of amalgamated datasets to provide access to their combined information [232].

The original vision of the SW is encapsulated in the well-known layered diagram shown in Figure 3.1. As can be seen, the development process of the SW is moving upward, with the RDF/OWL nexus in the middle. RDF as noted sits on top of XML, and the lowest level of all is that of the Uniform Resource Identifier (URI). In the next subsection we examine the foundational role that the URI plays in the SW vision.

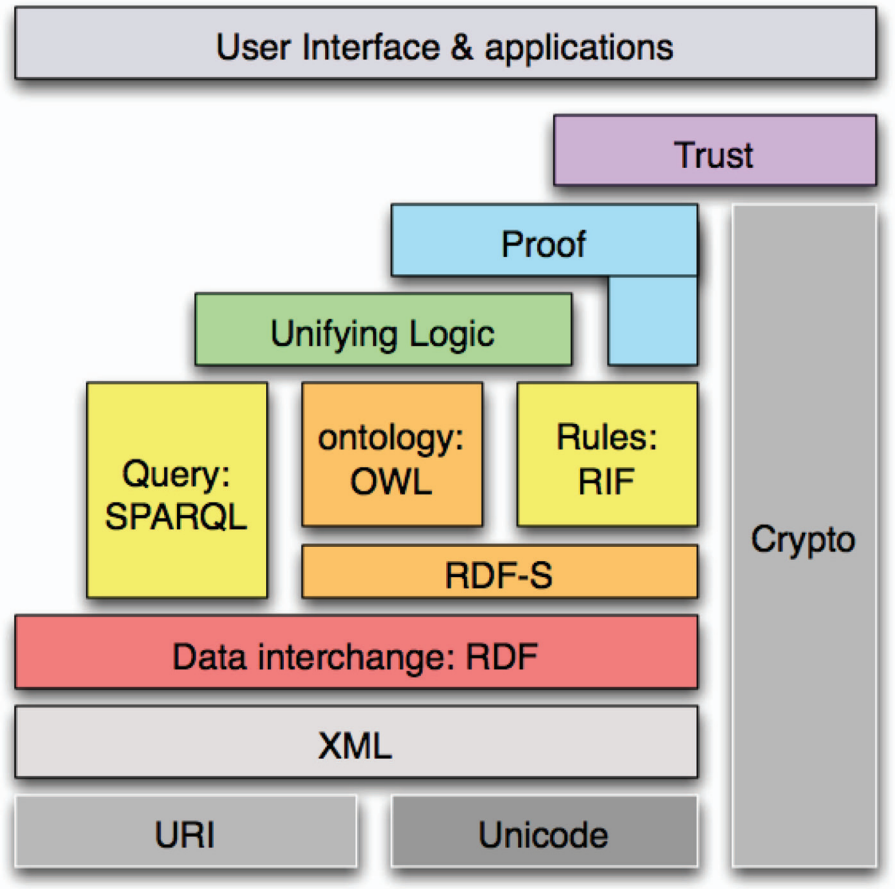

Fig. 3.2 The Semantic Web Stack c.2006.

The vision has moved on with the implementation effort, as one might expect. Following the implementation of ontologies using OWL, attention switched to the rules layer and appropriate languages for expressing rules; current thinking suggests that the Rule Interchange Format (RIF) currently under development [112] should sit alongside OWL as another extension of RDF-S. These layers are covered by the query language SPARQL. This revised vision of the SW stack, together with recognition of the need for effective user interfaces and applications, is shown in Figure 3.2.

3.1.2 URIs: Names or addresses? Or both?

RDF is based on the identification of resources via URIs, and describing them in terms of their properties and property values [195]. Compare RDF with XLink, the linking language for XML, which provides some information about a link but doesn't provide any external referent to anything with respect to which the link is relevant. In contrast, RDF assigns specific URIs to individual things, as we see in the following example. As we create the RDF graph of nodes and arcs (Figure 3.3), we can see that URIs are even used for the relations. A URI reference used as a node in an RDF graph identifies what the node represents; a URI used as a predicate identifies a relationship between the things identified by the nodes that are connected [172].

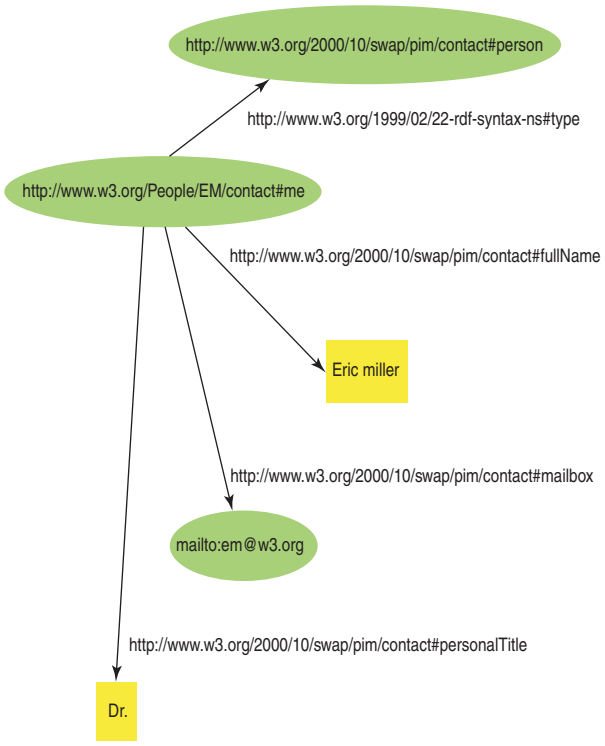

Fig. 3.3 RDF graph showing URIs.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
<contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
<contact:fullName>Eric Miller</contact:fullName>
<contact:mailbox rdf:resource="mailto:em@w3.org"/>
<contact:personalTitle>Dr.</contact:personalTitle>
</contact:Person>
</rdf:RDF>
```

In general, using URIs to identify resources is an important factor in the development of the Web [33]. Using a global naming syntax convention (however arbitrary *qua* syntax) provides global network effects, from which the benefits of the Web derive; URIs have global scope and are interpreted consistently across contexts. Associating a URI with a resource should happen if anyone might reasonably wish to link to it, refer to it or retrieve a representation of it [155].

Relations, identified by URIs, link resources which are also identified by URIs. To get the machine-readability that the SW is intended to secure, then the machine needs to be able to get at the relation, and therefore must be able to dereference the URI that identifies the relation and retrieve a representation of the identified resource. If the relevant information about the relation (for example, property restrictions) is also available at the URI, then the machine will be able to make inferences about the relation asserted. RDFS and the more complex OWL allow the assertion of property restrictions which in turn allows the machine to make inferences in this way. In this way, the SW is underpinned by URIs; the use of URIs allows machines to process data directly enabling the intended shift of emphasis from documents to data. We noted above that much of the inspiration for the SW comes from relational databases; in order to achieve the anticipated gains in functionality with respect to a particular database, the objects in the database must be exported as first class objects to the Web, and therefore need to be mapped into a system of URIs. The linking that underpins the SW is of course intended to provide a generic infrastructure for machine-processable Web content, but it has been argued that this infrastructure also addresses many of the concerns of the traditional hypermedia community [278].

Performing this foundational function needs a shift in our understanding of how we use URIs. Typically, names and addresses are different; the name of something refers directly to it, the address tells you where it is (if not exactly how to get hold of it). In traditional computing identifiers turned up in programming languages, addresses were locations in memory. Names are nailed to objects, addresses to places, and therefore an object should have one name forever while its address may change arbitrarily often. This in some ways fed into a "classical" view of the Web: there was an assumption that an identifier (a URI) would be one of two kinds of thing. It would either be the name of something, understood separately from location – a URN – or specify the location of that thing – a URL. So the class of URIs partitioned into the class of URNs and the class of URLs (and maybe one or two others, such as Uniform Resource Citations). The HTTP scheme, for example, was seen as a URL scheme.

This extra layer of conceptual complication was gradually seen to be of less use, and the notion of a URI became primary. URIs can do their identifying either directly or via location, but this is not a deep conceptual distinction. Hence HTTP is a URI scheme, although an HTTP URI identifies its object by representing its primary access mechanism, and so (informally) we can talk about the HTTP URI being a URL. The name/address distinction is a spatial metaphor that works perfectly well in a standard computing environment, but in networked computing systems the distinction breaks down. Similarly, objects can be renamed, and often are (reasons why they shouldn't be are discussed in Section 5.4.6 below). If a hierarchical system of naming is set up and maintained by an authority, then the name will function only as long as that authority supports that hierarchical system, and at the limit only as long as the authority itself remains in existence.

So we should beware of pressing the analogy of the spatial name/address system too closely. A literal location is a point in 3-D space, and within networked computer systems we should not get too fixed on what we should call names, or addresses, or the physical location of the memory cell that will store them. A computer memory address is often an address in a virtual memory space allocated to an object, which is translated in use by hardware into a physical memory address. IP addresses aren't bound to particular computers,

but implicitly contain reference to routing information, so the computer corresponding to a given IP address cannot be moved far in the routing structure. Domain names get used to refer to a computer or what the computer presents when we wish to reserve the right to move the thing corresponding to the identification from one part of the Internet to another. So the Domain Name System (DNS), being independent of the routing system, does not restrict the IP addresses that can be given to a computer of a given domain name. DNS does look like a system of names, whereas IP addresses do seem to function like addresses [26].

However, it is also very observable that domain names for particular resources do change, because the protocols used for naming them are altered – the reason being that there is information embedded in the name. In the offline world, names can survive the failure of such embedded information to remain true (John Stuart Mill gives the example of 'Dartmouth' as a place whose location may or may not remain at the mouth of the River Dart). Such changes are unproblematic. But online, this is harder to ensure.

Consider the example http://pegasus.cs.example.edu/disk1/ students/ romeo/cool/latest/readthis.html [26]. There are all sorts of reasons why this URI might change. 'pegasus', 'cs', 'students' etc may all change over the years as different computers get used to host the information, or as Romeo graduates and becomes a faculty member. His opinions about what is 'cool' or what is 'latest' will also evolve over time (one hopes). 'http', being the protocol used to present the resource, and 'readthis' being relatively meaningless are the least likely parts of the URI associated with the particular resource to change.

The reason such information is included is because a name has to be dereferenced in order to find out anything about what the name is naming. Typically that involves using some sort of index or set of indexes, which may be official and canonical, or informal and unofficial, to look up the name. Such indexes are often hierarchical to facilitate lookup, as DNS names are. It would be possible to omit all information from a domain name, and ensure a unique identifier for the resource (and indeed there would then be no obvious reason, all things being equal, why the identifier shouldn't be permanent as well), at the cost of making it hard to look up and dereference.

Such matters were of relatively small significance as long as humans were the main users and exploiters of the Web – after all, one is mainly after a resource and the content it contains, and although it may be frustrating to follow a URI only to find the resource no longer lived there, that was an irritation rather than a serious breakdown in the system. People are also relatively flexible in online retrieval and can tolerate ambiguities. But some kind of resolution to the name/address issue is required if we expect formal systems to deal with URIs. The SW is a tool for doing things in a social space, not merely a set of rules for manipulating formulae, so we need to know what we are referring to, and how to get at those referents where appropriate. It is desirable for an e-commerce system, for example, to refer without ambiguity to a number of things: documents such as bills and invoices, abstract items such as prices, and concrete things like buyers and the items that are actually bought and sold. [31] summarises and provides a critique of a large number of ways of understanding this issue in the context of HTTP.

Naming, ultimately, is a social set of contractual arrangements. We should not let the virtual nature of cyberspace blind us to the fact that people ask and pay for, and get given, domain names and space on servers. Authorities maintain these things, and also act as roots for dereferencing purposes. The stability of these institutional setups will help determine the stability of the Web's naming system.

3.1.3 Ontologies

Above RDF and RDFS in Figure 3.2 sits the ontology. On a traditional conception [123], ontologies contain specifications of the concepts that are needed to understand a domain, and the vocabulary required to enter into a discourse about it, and how those concepts and vocabulary are interrelated, how classes and instances and their properties are defined, described and referred to. An ontology can be formal or informal. The advantage of formality is that it makes the ontology machine-readable, and therefore allows a machine to undertake deeper reasoning over Web resources. The disadvantage is that such formal constructs are perceived to be hard to create.

Data can be mapped onto an ontology, using it as a *lingua franca* to facilitate sharing. Ontologies are therefore intended to put some sort of order onto information in heterogeneous formats and representations, and so contribute to the ideal of seeing the Web as a single knowledge source. To that extent, an ontology is similar to a database schema, except that it will be written with a comparatively rich and expressive language, the information will be less structured, and it determines a theory of a domain, not simply the structure of a data container [96].

So ontologies are seen as vital adjuncts to data sharing, and the ultimate goal of treating the Web as a single source of information, but they also have detractors. Many commentators worry that the focus on ontologies when it comes to postulating formalisms for the future Web is to make the mistake of over-privileging classification when it comes to understanding human language and communication [113]. It should certainly be pointed out that many ontologies actually in use, for example in industry, are taxonomies for special-purpose classification of documents or webpages, tend not to be elaborate, and do not rely on highly expressive formalisms [88].

OWL has its roots in an earlier language DAML+OIL [65] which included description logic (DL – [42]) among its various influences. OWL goes beyond DL, which sets out domain concepts and terminology in a structured fashion, by using the linking provided by RDF to allow ontologies to be distributed across systems, compatible with Web standards, open, extensible and scalable. Ontologies can become distributed as OWL allows ontologies to refer to terms in other ontologies. In this way OWL is specifically engineered for the Web and Semantic Web, and of many languages sharing symbols ([cf. 134]).

It is difficult to specify a formalism that will capture *all* the knowledge, of arbitrary type, in a particular domain. Ontologies, of course, serve different purposes, and can be deeper (expressing the scientific consensus in a discipline, and correspondingly labour-intensive to construct) or more shallow (with relatively few terms that organise large amounts of data – [34]). Indeed, there are many other types of discourse beyond ontologies of course, and many logics for expressing them, for example causal, temporal and probabilistic logic.

Causal logic [e.g. 258] developed out of logics of action in AI, and is intended to capture an important aspect of common sense understanding of mechanisms and physical systems. Temporal logic formalises the rules for reasoning with propositions indexed to particular times; in the context of the fast-growing Web, the prevalence of time-stamping online and the risks of information being used that is out of date ensures the relevance of that. Certainly temporal logic approaches have been suggested for ontology version management [149].

Probabilistic logics are calculi that manipulate conjunctions of probabilities of individual events or states, of which perhaps the most well-known are Bayesian, which can be used to derive probabilities for events based on prior theories about how probabilities are distributed (and very limited real data). Bayesian reasoning is commonplace in search engines, and even the search for spam (cf. [117]). In domains where reasoning under uncertainty is essential, such as bioinformatics, Bayesian ontologies have been suggested to support the extension of the Web to include such reasoning [19]. The utility of Bayesian approaches in computational systems cannot be doubted; more controversially some also claim that human reasoning conforms to a Bayesian pattern [118], although a significant body of work suggests humans are not Bayesian estimators [162]. Notwithstanding, at the very least machines that consistently adjust their probabilities in the light of experience will have a complementary role supporting human decision making.

The Web is often falsely conceived as being static, whereas it is constantly changing. Dynamic semantics relate to the activities surrounding the content of the Web: creating content, user-guided action, time, users' personal profiles and so on [104]. Fry et al, who are supporters of the SW project, argue that the assumptions underlying the vision of the SW are that semantics are declarative – we are dealing with passive data that can be retrieved from a server – and that changes are slow – publishing events are much rarer than browsing or clicking on a link. On the other hand, the context of retrieval, such as the user's profile and what tasks he or she is engaged in at retrieval time, is also an issue, as is the browsing context (different patterns of navigation may mean different sets of relations and informational contexts need to be understood), agents dynamically computing metadata, or the usual process of editing the Web creating different editions of a page.

Hence there is certainly logical and conceptual apparatus that will enable a rich variety of reasoning to be expressed, although the deeper argument made by many critics, such as [113], that a great many limitations result from the situated, embodied and embedded nature of much reasoning and conceptualisation, won't be addressed by the proliferation of abstract formalisms. But equally we must try to avoid the assumption that the SW is intended as a single overarching system, with a single way of interacting and a particular set of representation requirements that force all knowledge into one form (cf. [158]).

As we have seen, the SW is intended in particular to exploit one type of data, relational data. If such data have value in a context, then SW technologies should similarly have value, and indeed should add value as they should (a) enable further inference to be done on the data, and (b) allow, via ontologies, the data to be linked to potentially vast stores of data elsewhere. The SW claim, then, is not that all data or knowledge has to be representable in some narrow set of formalisms, but rather that the power of linking data allows much more to be done with it. For many purposes, and in some contexts for most usual purposes, unambitious representation schemes that may appear to lack a rich range of expressive possibilities may well be entirely adequate. The SW is not intended to be a system that will meet all purposes, but it is an extension of the Web that is intended to exploit the potential of the linking of unprecedented quantities of data. Ontologies will allow a common understanding of data pulled together from heterogeneous sources, as long as their relevant parts are appropriate for the task in hand. The ambition is in the range of data that such an approach can exploit, and in the value SW technologies hope to add, not in the extension of the range of inference that can be achieved automatically (though extending the range should also be possible).

3.1.4 Folksonomies and emergent social structures

The use of ontologies adds structure to data. However, structure can emerge organically from individuals' management of their own information requirements, as long as there are enough individuals. There are increasingly many applications driven by decentralised communities from the bottom-up, which go under the ill-defined but

popular name of *social software*. For instance, a *wiki* is a website that allows users and readers to add and edit content, which allows communication, argument and commentary; the Wikipedia (http://en.wikipedia.org/wiki/Main_Page for the English language version), an online encyclopaedia written by its user community, has become very reliable despite ongoing worries about the trustworthiness of its entries and fears of vandalism. Ontologies may be supplemented by *folksonomies*, which arise when a large number of people are interested in some information, and are encouraged to describe it – or *tag* it (they may tag selfishly, to organise their own retrieval of content, or altruistically to help others' navigation). Rather than a centralised form of classification, users can assign key words to documents or other information sources. And when these tags are aggregated, the results are very interesting. Examples of applications that have managed to harness and exploit tagging are Flickr (http://www.flickr.com/ – a photography publication and sharing site) and del.icio.us (http://del.icio.us/ – a site for sharing bookmarks). Keepers of unofficial weblogs (*blogs*) tag their output. The British Broadcasting Corporation (BBC) has seen opportunities here with a radio programme driven by users' tagging (via mobile phone) of pop songs [61].

As the number of tags on an application increases, increasing structure is detectable – tags tend to be reused, and reapplied to new items by new users, and all the usual relationships of subsumption, etc, start to emerge. The resulting rough structures are folksonomies (= folk taxonomies). They are certainly illogical and idiosyncratic, and contain many confusing examples of synonymy (several words meaning the same thing – science fiction, sci-fi and SF) and polysemy (several meanings covered by the same word – does SF = science fiction or San Francisco?), which will hinder efficient search – and of course are language-dependent. Not only that, but one imagines that as tag structures are used increasingly frequently to organise certain Web applications, the spammers will start tagging automatically to boost the chances of their data being retrieved. On the other hand, tags are generated by real-world interaction with the tagged content, and so do reveal genuine patterns of engagement between the content providers and users. The evolution of tags, over very large sets of

tagging data, can be tracked to show these patterns developing through time [84].

Such structures allow semantics to *emerge* from implicit agreements, as opposed to the *construction* of ontologies indicating *explicit* agreements; the field of *semiotic dynamics* is premised on the idea that agreed communication or information organisation systems often develop through similar decentralised processes of invention and negotiation [268]. It has been argued that implicit agreement, in the form of on-demand translations across information schemas can be adequate to support interoperable semantics for, and distributed search through, P2P systems – though whether such implicit translations will be easy to generate across information sources designed for different tasks is very much an open question [2].

3.1.5 Ontologies v folksonomies?

It is argued – though currently the arguments are filtering only slowly into the academic literature – that folksonomies are preferable to the use of controlled, centralised ontologies [e.g. 259]. Annotating Web pages using controlled vocabularies will improve the chances of one's page turning up on the 'right' Web searches, but on the other hand the large heterogeneous user base of the Web is unlikely to contain many people (or organisations) willing to adopt or maintain a complex ontology. Using an ontology involves buying into a particular way of carving up the world, and creating an ontology requires investment into methodologies and languages, whereas tagging is informal and quick. One's tags may be unhelpful or inaccurate, and no doubt there is an art to successful tagging, but one gets results (and feedback) as one learns; ontologies, on the other hand, require something of an investment of time and resources, with feedback coming more slowly. And, crucially, the tools to lower the barriers to entry to controlled vocabularies are emerging much more slowly than those being used to support social software [61].

Tagging is certainly an exciting development and an interesting phenomenon, but we should be wary of assuming that tags and ontologies are competing for the same space. Tagging provides a potential source

of metadata, with all the disadvantages of informality and all the advantages of low barriers to entry and a high user base. But tags are only part of the story about Web resources [128].

Ontologies and folksonomies have been caricatured as opposites. In actual fact, they are two separate things, although some of the functionality of ontologies can uncontroversially be taken over by folksonomies in a number of contexts. There are two separate (groups of) points to make. The first has to do with the supposed trade-off between ontologies and folksonomies; the second to do with perceptions of ontologies.

Ontologies and folksonomies are there to do different things, and deal with different cases. Folksonomies are a variant on the keyword-search theme, and are an interesting emergent attempt at information retrieval – how can I retrieve documents (photographs, say) relevant to the concept in which I am interested? Ontologies are attempts to regulate parts of the world of data, and to allow mappings and interactions between data held in disparate formats or locations, or which has been collected by different organisations under different assumptions. What has been represented as a trade-off, or a competition, or even a zero-sum game may be better represented as two separate approaches to two different types of problem. It may be that the sets of problems they are approaches to overlap, in which case there may on occasion be a choice to be made between them, but even so both ontologies and folksonomies have definite uses and are both potentially fruitful avenues of research [257].

It has been argued that ontologies could usefully incorporate material from social networks and software, as the information being modelled has a social dimension [201]. This may offer a new set of opportunities – for example blogging software that automatically creates metadata could be a way to exploit the bottom up social software approach [163]. Furthermore, the supposed basis of the distinction between the two – that folksonomies evolve organically and painlessly whereas ontologies are high maintenance and high overhead – is anyway dubious. Where there is a perceived need for ontologies, lightweight but powerful ones do emerge and are widely used, as for instance with Friend-of-a-Friend (FOAF – [45]), and associated applications such as Flink [200]. This fits in general with calls for the dual and

complementary development of SW technologies and technologies that exploit the self-organisation of the Web [e.g. 101].

Perceptions of ontologies depend on the understanding of this distinction. Consider, for example, the costs of ontologies. In the first place, there will be areas where the costs, be they ever so large, will be easy to recoup. In well-structured areas such as scientific applications, the effort to create canonical specifications of vocabulary will often be worth the gain, and probably essential; indeed, Semantic Web techniques are gaining ground in scientific contexts with rich data in which there exists the need for data processing and the willingness to reach a consensus about terms. In certain commercial applications, the potential profit from the use of well-structured and coordinated specifications of vocabulary will outweigh the sunk costs of developing or applying an ontology, and the marginal costs of maintenance. For instance, facilitating the matching of terms in a retailer's inventory with those of a purchasing agent will be advantageous to both sides.

And the costs of developing ontologies may decrease as the user base of an ontology increases. If we assume that the costs of building ontologies are spread across user communities, the number of ontology engineers required increases as the log of the size of the user community, and the amount of building time increases as the square of the number of engineers – simple assumptions of course but reasonable for a basic model – the effort involved *per user* in building ontologies for large communities gets very small very quickly [29]. Furthermore, as the use of ontologies spreads, techniques for their reuse, segmentation and merging will also become more familiar [212, 256, 10], and indeed there will be an increasing and increasingly well-known base of ontologies there to be reused.

Secondly, there is a perception of ontologies as top-down and somewhat authoritarian constructs, unrelated, or only tenuously related, to people's actual practice, to the variety of potential tasks in a domain, or to the operation of context (cf. e.g. [158]). In some respects, this perception may be related to the idea of the development of a single consistent Ontology of Everything, as for example with CYC [183]. Such a wide-ranging and all-encompassing ontology may well have a number of interesting applications, but clearly will not scale and its

use cannot be enforced. If the SW is seen as requiring widespread buy-in to a particular point of view, then it is understandable that emergent structures like folksonomies begin to seem more attractive (cf. [259]).

But this is not an SW requirement. In fact, the SW's attitude to ontologies is no more than a rationalisation of actual data-sharing practice. Applications can and do interact without achieving or attempting to achieve global consistency and coverage. A system that presents a retailer's wares to customers will harvest information from suppliers' databases (themselves likely to use heterogeneous formats) and map it onto the retailer's preferred data format for re-presentation. Automatic tax return software takes bank data, in the bank's preferred format, and maps them onto the tax form. There is no requirement for global ontologies here. There isn't even a requirement for agreement or global translations between the specific ontologies being used *except* in the subset of terms relevant for the particular transaction. Agreement need only be local.

The aim of the SW should be seen in the context of the routine nature of this type of agreement. The SW is intended to create and manage standards for opening up and making routine this partial agreement in data formats; such standards should make it possible for the exploitation of relational data on a global scale, with the concomitant leverage that that scale buys.

3.1.6 Metadata

The issues pertaining to the semantics or interpretation of the Web go beyond the Semantic Web. For instance, metadata can be used to describe or annotate a resource in order to make it (more) intelligible for users. These users might be human, in which case the metadata can be unstructured, or machines, in which case the metadata have to be machine-readable. Typically, metadata are descriptive, including such basic elements as the author name, title or abstract of a document, and administrative information such as file types, access rights, IPR states, dates, version numbers and so on. Multimedia items may be annotated with textual descriptions of the content, or key words to aid text-based search.

In general, metadata are important for effective search (they allow resources to be discovered by a wide range of criteria, and are helpful in adding searchable structure to non-text resources), organising resources (for instance, allowing portals to assemble composite webpages automatically from several suitably-annotated resources), archiving guidance (cf. [58]), and identifying information (such as a unique reference number, which helps solve the problem of when one Web resource is the 'same' as another). Perhaps the most important use for metadata is to promote interoperability, allowing the combination of heterogeneous resources across platforms without loss of content. Metadata schema facilitate the creation of metadata in standardised formats, for maximising interoperability, and there are a number of such schemes, including the Dublin Core (http://dublincore.org/) and the Text Encoding Initiative (TEI – http://www.tei-c.org/). RDF provides mechanisms for integrating such metadata schemes.

There are a number of interesting questions relating to metadata. In the first place, what metadata need to be applied to content? Secondly, how will metadescription affect inference? Will it make it harder? What can be done about annotating legacy content? Much has been written about all these questions, but it is worth a small digression to look at some approaches to the first.

As regards the metadata required, needless to say much depends on the purposes for which resources are annotated. For many purposes – for example, sharing digital photos – the metadata can look after themselves, as the success of sites like Flickr show. More generally, interesting possibilities for metadata include time-stamping, provenance, uncertainty and licensing restrictions.

Time-stamping is of interest because the temporal element of context is essential for understanding a text (to take an obvious example, when reading a paper on global geopolitics in 2006 it is essential to know whether it was written before or after 11th September, 2001). Furthermore, some information has a 'sell-by date': after a certain point it may become unreliable. Often this point isn't predictable exactly, but broad indications can be given; naturally much depends on whether the information is being used in some mission critical system and how tolerant of failure the system is. General temporal information about a resource

can be given in XML tags in the normal way. However, in the body of the resource, which we cannot assume to be structured, there may be a need for temporal information too, for users to find manually. In such a case, it is hard to identify necessary temporal information in a body of unstructured text, and to determine whether a time stamp refers to its own section or to some other part of the resource. It may be that some ideas can be imported from the temporal organisation of more structured resources such as databases, as long as over-prescription is avoided [173]. In any case, it is essential to know the time of creation and the assumptions about longevity underlying information quality; if the content of a resource 'is subject to change or withdrawal without notice, then its integrity may be compromised and its value as a cultural record severely diminished' [107].

Provenance information is extremely important for determining the value and integrity of a resource. Many digital archiving standards set out clearly what provenance information are required. For instance, the Open Archival Information System model (OAIS) of the Consultative Committee on Space Data Systems demands metadata about the source or origin of a resource, a log of the changes that have taken place, and under whose aegis, and a record of the chain of custody [57]. The CURL Exemplars in Digital Archives project (Cedars) goes further, demanding a history of origins (including the reasons why the resource was created, the complete list of responsible custodians since creation and the reason it is being proposed for archiving), technical information about the creation environment of the document (including software and operating systems), management history (including the history of archiving process and the policies and actions applied to it since it was archived), and a record of the IPR relating to the document [58]. Technological contexts such as e-science and grid computing have prompted research into the technology-independent representation of provenance, the provenance information that needs to be encoded, key roles for a provenance-recording architecture and process-related items such as an architecture's distribution and security requirements (cf. [122] – ironically a currently evolving document at the time of writing that includes an unstructured account of its own provenance).

Another key factor in assessing the trustworthiness of a document is the reliability or otherwise of the claims expressed within it; metadata about provenance will no doubt help in such judgements but need not necessarily resolve them. Representing confidence in reliability has always been difficult in epistemic logics. In the context of knowledge representation approaches include: subjective logic, which represents an opinion as a real-valued triple (belief, disbelief, uncertainty) where the three items add up to 1 [159, 160]; grading based on qualitative judgements, although such qualitative grades can be given numerical interpretations and then reasoned about mathematically [110, 115]; fuzzy logic (cf. [248]); and probability [148]. Again we see the trade-off that the formalisms that are most expressive are probably the most difficult to use.

Finally, metadata relating to licensing restrictions has been growing with the movement for 'creative commons', flexible protections based on copyright that are more appropriate for the Web and weblike contexts. Rather than just use the blunt instrument of copyright law, creative commons licenses allow authors to fine-tune the exercise of their rights by waiving some of them to facilitate the use of their work in various specifiable contexts [187]. We discuss copyright in more detail in Section 6.2 below.

The questions about the difficulties of reasoning with metadata, and the giant task of annotating legacy data, remain very open. It has been argued that annotating the Web will require large-scale automatic methods, and such methods will in turn require particular strong knowledge modelling commitments [170]; whether this will contravene the decentralised spirit of the Web is as yet unclear. Much will depend on creative approaches such as annotating on the fly as annotations are required, or else annotating legacy resources such as databases underlying the deep Web [283].

3.2 Reference and identity

The Semantic Web relies on naming conventions with URIs, and of course every part of the Web's labelling system relies on some convention or other. The problem with labelling on the Web is that any

system is essentially decentralised and not policed, in accordance with the Web's governing principles, but this lack of centralisation allows different schemes and conventions, and indeed carelessness, to flourish, which in turn opens up the possibility of failures of unique reference.

3.2.1 Reference: When are two objects the same?

Decentralisation is a problem from a logical point of view, though a huge advantage from that of the creator of content. The same object might be referred to online, perfectly correctly, as 'Jane Doe', 'Janey Doe', 'Jane A. Doe', 'Doe, J.A.' and so on. Furthermore, any or all of these terms may be used to refer to another distinct object. And, of course, the original Jane Doe may be misnamed or misspelled: 'Jnae Doe', etc. These failures of unique reference are relatively trivial for human users to disentangle, but are of course very hard for machines to work out. And if we are hoping to extract usable information from very large repositories of information, where hand-crafted solutions and checking reference by eye are not feasible, machine processing is inevitable. Reference problems are particularly likely to occur when information sources are amalgamated, a ubiquitous problem but a serious one in the context of the Semantic Web. And the decentralisation of the Web precludes making a unique name assumption, in the manner of [240].

On the other hand, URIs provide the Web with the resources to avoid at least some traditional grounding problems, when it can be resolved that two terms are pointing to the same URI. So if "morning star" and "evening star" both point directly to http://ex.org/planets.owl#venus then any further grounding is superfluous. On the other hand, two different URIs might refer to the same object non-obviously, and may do so through only some operations in which it is used. Sometimes this will be detectable through algorithmic analysis of syntax (for example, domain names are not case sensitive, so this could be used to detect similarity), but not in general. The problem doesn't go away with the use of URIs, but they are at least a set of identifiers giving a potential basis for stability in some situations – particularly scientific situations where agreement over symbols and definitions is often formalised.

A heuristic method for resolving such clashes, in the real world, is to make an intelligent judgement based on collateral information, and this has been mimicked online by the computation of the community of practice of a name, based on the network of relationships surrounding each of the disputed instances. For example, if 'Jane Doe' and 'Doe, J.A.' have both got strong associations with 'University of Loamshire', one because she works there, the other because she has worked on a project of which UoL was a partner, then that is *prima facie* evidence that the two terms refer to the same object – though of course such a judgement will always be highly defeasible [11].

In general, reference management, and the resolution of reference problems, will always be tricky given that the Web covers a vast amount of information put together for a number of different reasons and to solve various tasks; meanings and interpretations often shift, and there may on occasion be little agreement about the referents of terms. An important issue for Web Science is precisely how to understand reference and representation, and to determine which management systems and formalisms will allow greater understanding and tracking of what the Web is purporting to say about which objects.

3.2.2 When are two pages the same?

An alternative take on the reference problem is that of determining when two webpages are *the same page*. This of course would be trivial in many cases, but often the "main" text is copied from one page to another, but surrounded by different advertisements, logos, headers and footers. Many metrics are available that are intended to determine quantitatively the extent of the relation between two pages. Similarity judgements can be arbitrary and pragmatic, depending on context (e.g. deciding plagiarism or copyright infringement cases), but techniques from information theory do exist to produce objective sets of numbers to feed into the judgement process – for instance, the Levenshtein edit distance, and variant algorithms, given by the minimum number of operations from some base set needed to transform one string into another (cf. [38]).

The basis for making similarity judgements need not only be the content on the page, but could also be the structure of hyperlinks within which the page is embedded. The information that a user requires need not come from a single page, but instead can be gleaned from the cluster of documents around the basic topic, and so the linkage structures there can be extremely important. And a further possible way of understanding similarity is between particular usage patterns of the page – are two pages often accessed at similar points in a Web surfing session [76]?

Content-based similarity can be approached by matching words or subsequences from the two pages. Relatively simple techniques can be used to determine the resemblance between two pages (the ratio between the size of the intersection of the subsequences and the size of their union), and the containment of one by the other (the ratio between the intersection and the size of the contained set) [48]. Link-based metrics derive from bibliometrics and citation analysis, and focus on the links out and links in that two pages have in common, relative to the general space of links in the topic cluster. Usage-based metrics exploit information gathered from server logs and other sources about when pages are visited, on the assumption that visits from the same user in the same session in the same site are likely to be conceptually related, and the greater the similarity between the times of users' access to webpages, the greater the likelihood of those pages being somehow linked conceptually [227].

3.3 Web engineering: New directions

The Web's development is a mix of standard-setting, unstructured, decentralised activity and innovation, and deliberate engineering. In this section we will focus on the latter, and review prominent engineering issues and open imperatives. The growth of the Web is clearly a key desideratum. The storage of ever-larger quantities of information, in the context of ever-quicker computation, will be vital for the foreseeable future. Without smarter storage and faster retrieval for memory-hungry media like video, then ultimately the Web will grow too large for its own technologies. For instance, PageRank requires crawling and caching of significant portions of the Web; Google's success depends

on being able to keep its cache tractable while also of a significant size. Greater demand for personalised services and search will also put pressure on the system. Widening the scope of search to encompass items such as multimedia, services or ontology components, will also require the pursuit of academic research programmes, effective interfaces and plausible business models before commercial services come on stream. Existing and developing approaches to leveraging the Web have to be extended into new Web environments as they are created (such as P2P networks, for example).

3.3.1 Web services

Services are a key area where our engineering models of the Web need to be engaged and extended. Web services are distributed pieces of code written to solve specific tasks, which can communicate with other services via messages. Larger-scale tasks can be analysed and recursively broken down into subtasks which with any luck will map onto the specific tasks that can be addressed by services. If that is the case, and if services are placed in a Web context, that means that users could invoke the services that jointly and cooperatively meet their needs.

Software abstracts away from hardware and enables us to specify computing machines in terms of logical functions, which facilitates the specification of problems and solutions in relatively intuitive ways. The evolution of the Web to include the provision and diffusion of services opens up new abstraction prospects: the question now is how we can perform the same abstraction away from software. What methods of describing services will enable us to cease to worry about how they will be performed?

A number of methods of specifying processes have developed over the last few years and applied to the Web service domain. For example, WS-Net is an architectural description language based on the theory of coloured Petri nets (i.e. an extension of simple Petri net theory with valued, identifiable tokens – see Section 4.2.5 for a brief discussion of Petri nets), which describes a Web service component in terms of the services it provides to other components, the services it requires to function, and its internal operations. The end result is a model that encompasses both the global and the local aspects of a service system,

facilitates Web service integration to achieve new goals, while also providing a formalism for integration evaluation [296].

Process algebras (see Section 4.2.5) have also been applied to services. Again, as with the Petri net approach, the use of a formal algebra allows both design and evaluation to take place (or indeed one or the other, depending on what alternative methods are available to generate or survey the code). For instance, [98] describes the mapping between an expressive process algebra and BPEL4WS (a standardised XML-based notation for describing executable business processes), which allows both the creation of services in BPEL4WS followed by their evaluation and verification using the process algebra, or the generation of BPEL4WS code automatically from the use of the algebra to specify the desired services. In general, the algebraic specification of services at an abstract level and reasoning about them has been a major area of research on services [e.g. 75, 105, 208].

BPEL4WS is an extended version of the Business Process Execution Language BPEL, which is becoming an increasingly important way to interleave Web services with business processes. BPEL has its limits, but allows the creation of composite services from existing services. The next stage is to adapt this approach for the P2P environment, and the vehicle currently under development for that is CDL, aka WS-CDL, aka Choreography (Web Services Choreography Description Language – [164]), an XML-based language for defining the common and complementary observable behaviour in P2P collaborations. The aim is that interoperable P2P collaborations can be composed using Choreography without regard to such specifics as the underlying platforms being used; instead the focus is on the common goal of the collaborators. Whereas BPEL allows existing services to be combined together, Choreography shifts the focus onto the global description of collaborations, information exchanges, ordering of actions and so on, to achieve agreed goals.

3.3.2 Distributed approaches: Pervasive computing, P2P and grids

There are many hardware environments which the Web will be expected to penetrate, yet where engineering assumptions that apply

to large-scale, more-or-less fixed dedicated computing machines don't necessarily apply. Obvious examples include mobile computing, ubiquitous (or pervasive) computing where interoperability becomes an issue, P2P systems and grid computing. Mobile computing makes all sorts of engineering demands; the computing power available isn't vast and users must be assumed to be constantly on the move with variable bandwidth and access. Furthermore, presenting information to the user requires different paradigms from the PC, for example to allow the user to receive enough information on the small screen to make browsing compelling [20, 193]. Mobile access to the Web may become the dominant mode in many nations, particularly developing ones, thanks to relatively low prices and reliability of wireless connections and battery power [222]. Research in this area is important for the equitable distribution of Web resources.

Ubiquitous computing, P2P and grid computing share many serious research issues, most notably the coordination of behaviour in large scale distributed systems. Ubiquitous computing envisages small, relatively low-powered computing devices embedded in the environment interacting pervasively with people. There are various imaginative possibilities, such as smart threads which can be woven into clothing. But without second-guessing the trends it is clear that smaller devices will need wireless connections to network architectures allowing automatic *ad hoc* configuration, and there are a number of engineering difficulties associated with that issue (cf. [244, 176]).

For instance, service discovery in the pervasive paradigm must take place without a human in the loop. Services must be able to advertise themselves to facilitate discovery. Standards for publishing services would be required to ensure security and privacy, trust of the service's reliability, the compensation for the service provider, and exactly how the service would be composed with other invoked services to achieve some goal or solve the problem at hand [179].

This is just one example of a currently evolving computing environment that is likely to grow in importance. In the context of Web Science and the search for and description of the invariants of the Web experience, it is essential that the assumptions we make about environments, and the technologies that live in them, are minimised.

P2P networks, characterised by autonomy from central servers, intermittent connectivity and the opportunistic use of resources [220], are another intriguing environment for the next generation Web. In such networks (including file-sharing networks such as Napster, communication networks such as Skype, and computation networks such as SETI@home), the computer becomes a component in a distributed system, and might be doing all sorts of things: backing up others' files, storing encrypted fragments of files, doing processing for large-scale endeavours in the background, and so on. There are clearly many potential applications for both structured and unstructured P2P networks in the Web context. The question for Web scientists is what essential functions for the Web experience can be preserved in loosely coupled autonomous systems. Given the unusual characteristics of P2P, including the potentially great number and heterogeneity of P2P nodes, traditional engineering methods such as online experimentation (which would require unfeasibly large numbers of users to sign up to an architecture and allow their transactions to be monitored) or large-scale simulation (the scale is simply too large) will be inappropriate. The scale licensed by the Web, which we will continue to see in P2P networks, makes network modelling theory essential (cf. e.g. [249, 189]), but we must expect radical experimentation, innovation and entrepreneurialism to lead the way in this field.

The temptation to exploit radically decentralised environments such as P2P networks in the next generation of the Web is strong; decentralisation is a key aspect of the Web's success. So, for example, one could imagine P2P networks being used to locate cached pages for backups in the event of failure or error leading to missing pages or dangling links. It needs to be established whether the ability of a P2P network to do that (which itself is currently unproven) would undermine the domain name system or support it.

Whereas P2P systems exploit large scale distribution to achieve lots of small ends, grid computing [102] is often a distributed approach to large scale problems using substantial computing power to analyse enormous quantities of data. The problem is to coordinate the behaviour of a large number of computers, exploiting unused resources opportunistically like P2P; again like P2P, and unlike traditional distributed

computing, grid computing is meant to be neutral about administrative or platform boundaries. Open standards are therefore needed, and the Grid requires abstract descriptions of computing resources.

By analogy to the Semantic Web, the Grid has spawned the Semantic Grid, where information and computing resources are annotated with metadata (and as with the SW RDF is the language of choice), allowing the exploitation of machine-readable specifications for the automatic coordination of resources to solve particular large-scale problems [72]. The application of the Grid and the Semantic Grid to large scale problems shows enormous promise – indeed as data from CERN's Large Hadron Collider come on stream at a gigabyte/sec, the Grid is indispensable.

The Grid and Semantic Grid raise a number of old questions in a new guise. Given that one's computing resources are given over to outsiders, trust and security will require reconsideration [23]. Socially, an interesting issue is understanding whether the Grid will actually change science, or merely allow the processing of more and more data [207].

In general, all these new computing paradigms raise the question of how lots of relatively autonomous individuals can work together to produce mutually beneficial results (either results beneficial to each individual, or to society as a whole). Coordination problems such as these have always loomed large in many disciplines, and we shouldn't be surprised to find them at the centre of Web Science.

3.3.3 Personalisation

It has often been claimed that personalisation is important for leveraging the value of a network [81], and increasing consumer lock-in [281]. Allowing users to personalise their tools and workspace means that the Web remains more than a commoditised one-size-fits-all area and instead becomes a space within which people can carve out their own niches. Furthermore, they should also be able to receive better services, tailored to their own particular circumstances and preferences, for equal or only slightly more cost [90]. Recommender systems are an obvious application of the technology [6].

To get effective personalisation, there must be integrated use of information from a number of sources, including data about users (clickstream data, downloading patterns, online profiles), the resources being delivered (site content, site structure) and domain knowledge, together with data mining techniques sufficient to create a holistic view of the resources that includes as much of the information that users require, in a representation that will make sense for them, while excluding information they won't want, and which can take account of the dynamic nature of the user models. All that, while still preserving the relation between the invariants of the Web experience and the particular context of an individual's use that empower him or her to claim a corner of cyberspace and begin to use it as an extension of personal space. Given that, on the Web, the relevant information is likely to be highly distributed and dynamic, personalisation is expected to be one of the big gains of the Semantic Web, which is pre-eminently a structure that allows reasoning over multiple and distributed data sources.

There are many engineering programmes under way investigating heuristics for personalisation from the available information, including using machine learning [120], ontologies [74, 165], P2P networks [126], and producing representations to facilitate the gathering of user information [74, 157, 223], as well as providing environments that facilitate personalisation [136, 53, 194] and associative links based on user- rather than author-preferences [54]. Another important strand of personalisation engineering is the development of tools to enable relative neophytes to create or enhance complex knowledge engineering artefacts, such as ontologies [213, 211] or wrappers [250].

3.3.4 Multimedia

The Web is a multimedia environment, which makes for complex semantics – this is of course not a problem unique to the Web. Metareasoning and epistemology often presume a textual medium, even though actually much reasoning is in analogue form. For example experts often use diagrams to express their knowledge [174, 263]. There have been attempts to produce 'language-like' generative taxonomies of visual representations [190], but these have not seemed to have

interesting applications. Some researchers have tried to discover the principles that might underlie diagrammatic reasoning [60]. There have also been important applications for decoding of visual representations for the visually impaired [147] and visualising image collections against a domain ontology [8]. Ultimately, the integration of multimodal representations of the same scene or entity is a very hard problem [224]. In general, it is not known how to retrieve the semantics from non-textual representations reliably; this phenomenon is known as the *semantic gap*.

Nevertheless, the next generation Web should not be based on the false assumption that text is predominant and keyword-based search will be adequate for all reasonable purposes [127]. Indeed, the issues relating to navigation through multimedia repositories such as video archives and through the Web are not unrelated: both need information links to support browsing, and both need engines to support manual link traversal. However, the keyword approach may falter in the multimedia context because of the greater richness of many non-textual media [264]. The Google image search approach relies on the surrounding text for an image, for example, which allows relatively fast search, and again in general the user is often able to make the final choice by sifting through the recommendations presented (keyword-based image searches tend to produce many fewer hits, which may mean they are missing many plausible possibilities). The presence of the human in the loop is hard to avoid at the moment: human intervention in the process of integrating vision language with other modalities is usually required [224], although there are a number of interesting techniques for using structures generated from texts associated with image collections to aid retrieval in restricted contexts [7].

But it is always possible to expend more resources on analysing an image (say) in order to produce better matches for keyword searches, if speed is not an overriding factor [293]. In such feature analyses, an important issue is the relative importance of low-level features such as 'dominant colour', and high-level, abstract features or concepts, such as 'Madonna' or 'still life'. Search on low-level features may be speedier and more accurate, but users are likely to want quite abstract search terms [121].

As an interesting hybrid it has been suggested that the semantic gap could be filled with ontologies of the visual that include low-level terms and provide some sort of mapping onto higher-level abstract concepts expressed in queries and metadata [229]. Such an infrastructure has been created, using (i) a visual descriptors ontology based on an RDF representation of MPEG-7 visual descriptors, (ii) a multimedia structure ontology based on the MPEG-7 multimedia description scheme and (iii) a core ontology modelling primitives at the root of the concept hierarchy which is meant to act as the bridge between ontologies, all supplemented with a domain ontology [260]. A further important open issue is the interoperability of Semantic Web technologies with non-RDF-based metadata such as EXIF metadata on JPEGS or the informal image tags created in Flickr [279]. Further work is required on the relationship between human image retrieval requirements and the possibilities of automation [156, 206], including a deeper understanding of the relative capabilities of folksonomies and ontologies (see Sections 3.1.4–3.1.5).

Of course, the media here envisaged are image and video; open research questions remain not only about how far one could get in search by such an approach, but also about how many media will succumb to such an approach in an integrable way.

3.3.5 Natural language processing

Finally, there are substantial issues relating to natural language processing (NLP), the computational analysis of unstructured data in texts in order to produce machine understanding (at some level) of that text. NLP relates to the Web in a number of ways. In the first place, natural language is a very sparse domain, in that most sentences uttered or written occur once only or very rarely, and the giant scale of the Web provides a fascinating corpus for NLP reasoning. A recent guesstimate for the size of the Web was two thousand billion words, of which 71% were English, 6.8% Japanese and 5.1% German. Many relatively uncommon languages such as Slovenian or Malay can boast 100m words online, the same size as the widely-used and respected British National Corpus. There are arguments about how representative the Web is as

a corpus, but the notion of what a corpus should represent – should it include speech, writing, background language such as mumbling or talking in one's sleep, or errors for example? – is hard to pin down with any precision [167].

Secondly, given the problems of the Web scale, NLP techniques will be important in such tasks as summarisation (see, for instance, the annual Document Understanding Conference – http://duc.nist.gov/ and [69]), which may provide useful support for the human parts of the search task.

Thirdly, NLP has great potential for the construction of the sorts of intuitive interface that the heterogeneous and not necessarily computer-literate Web user community require. Indeed it may help bridge the gap between the SW vision of a Web made up of data manipulated logically, and the more traditional vision of the Web as a place where useful documents are retrieved. For instance, can NLP techniques be used to discover and express metadata [153]? Texts containing unstructured data can now be mapped onto existing resources such as ontologies to provide markup and annotation, after initial training sessions.

Computing ontologies such as we have encountered are different in purpose and structure from the thesauri and taxonomies from the NLP world, although there is a debate about the extent and nature of the distinction [125, 289]. WordNet, for example, is not an ontology strictly, for instance containing lexical items with different senses where an ontology tries to ensure a unique interpretation for the terms it uses. But equally WordNet does contain ontological relations like set inclusion and membership within it. NLP resources also have something in common with folksonomies and the like, as well as important differences.

From the point of view of Web Science, important open questions exist as to the relationship between NLP and the Web; are the statistical techniques used in NLP contrary or complementary to the logically and semantically based techniques of data interrogation used by the SW community? Or alternatively is there an optimal division of analytical labour between the two types of approach that we might exploit? Much depends on how we interpret the development of the Web. For instance, if one sees the major task as being to annotate and provide

rich context for content and structure ('taming the Web', as described in [196]), then NLP will play a key role in that, including mapping drift in meaning over time [290]. If we understand the Semantic Web as focusing on data and the relational database model, then logical terms and persistent URIs become central.

NLP works well statistically; the SW, in contrast, requires logic and doesn't yet make substantial use of statistics. Natural language is democratic, as expressed in the slogan 'meaning is use' (see Section 5.1 for more discussion of this). The equivalent in the SW of the words of natural language are logical terms, of which URIs are prominent. Thus we have an immediate disanalogy between NLP and the SW, which is that URIs, unlike words, have owners, and so can be regulated. That is not to say that such regulation will ensure immunity from the meaning drift that linguists detect, but may well provide sufficient stability over the short to medium term.

4

The analysis of the Web

Learning the properties of the Web as a formal object in its own right provides a good deal of leverage for designers of new systems, and even more perhaps for the standards bodies whose job it is to discover and preserve the essential invariants of the Web experience at the macro scale. In this section we will briefly review efforts to map the Web's topology, and then mathematical methods of investigation.

4.1 Web topology

4.1.1 The Web's structure

Topological investigations attempt to discern structure out of the basic elements of the architecture and the links between them. Structure can tell us a lot. The investigation of the structure of the Web is of course always dependent on the level of abstraction of its description. Such is the size of the Web that even very small differences in the performance of these components could make large differences at the macro level. For instance, though one would not generally be worried by the difference between an algorithm of $O(n)$ and an algorithm of $O(n \log n)$ in most problem spaces, on the Web scale the $\log n$ term could start to get appreciably large [191]. Hence the behaviour of the

components of large-scale networks is of relevance even when looking at the global properties of the Web.

Furthermore, structure in turn provides evidence of what conversations are taking place over the Web. Hence understanding structure is important for a number of applications, such as navigation, search, providing the resources to support online communities, or ameliorating the effects of sudden shifts in demand for information.

The Web is democratic to the extent that there is no centralisation or central coordination of linking. Conceived as a hypertext structure, its usability depends to a very large extent on effective linking; following a chain of badly linked pages leads to the well-known disorientation phenomenon of being 'lost in hyperspace'. Following a chain of links is also rendered less risky by Web browsers which contain 'back' buttons, which in effect provide the inverse of any hyperlink. And navigation need not only be a leisurely amble around a chain of hyperlinks, thanks to search engines that find pages with characteristics of interest to the user.

Web topology contains more complexity than simple linear chains. In this section, we will discuss attempts to measure the global structure of the Web, and how individual webpages fit into that context. Are there interesting representations that define or suggest important properties? For example, might it be possible to map knowledge on the Web? Such a map might allow the possibility of understanding online communities, or to engage in '*plume tracing*' – following a meme, or idea, or rumour, or factoid, or theory, from germination to fruition, or vice versa, by tracing the way it appears in various pages and their links [5]. Given such maps, one could imagine spotting problems such as Slashdot surges (the slowing down or closing of a website after a new and large population of users follow links to it from a popular website, as has happened from the site of the online magazine Slashdot) before they happen – or at least being able to intervene quickly enough to restore normal or acceptable service soon afterwards. Indeed, we might even discover whether the effects of Slashdot surges have declined thanks to the constant expansion of the Web, as has been argued recently [166].

Much writing about the Web seems to suggest that it is, in some ways, alive, evolving and out of control [e.g. 87], and the decentralised

model of the Web certainly promotes the view that its growth is beyond control. The Web-as-platform model means that there are genuine and powerful senses in which the "creators" of the Web (who can be conceived as: the early conceptualisers of pervasive links between knowledge and knowledge representations; the originators of the powerful standards and languages underlying the Web as we know it; the many professionals currently and selflessly undertaking the painstaking negotiations on W3C standards bodies; or the writers of the actual content that we see online) do not control the macroscopic structure. This model is very powerful, but that does not mean that the Web has necessarily become an undifferentiated soup of connected pages.

Methods of analysing the web looking at patterns of links [171] have turned out to be remarkably interesting, illuminating and powerful in the structures they uncover. For instance, some sites seem to be taken as *authoritative* in some way – in other words, many other sites link into them. Other sites contain many links out – one way of conceiving this would be that such sites index authorities on some topic – and these useful sites act as *hubs*. Such hubs may also be authorities, but equally they may be pointed to by few pages or even no pages at all. When methods such as those pioneered by Kleinberg, Brin and Page take the link matrix of the Web and find the eigenvectors, it turns out that they correspond to clusters around the concepts that the pages are about. Such authority-hub structures are of immense importance to our understanding of the Web, and require analysis of the link matrix to find. Indeed, Kleinberg's original intention was to discover authorities, and the ubiquity online of the more complex authority-hub structure was initially a surprise [171].

Several authorities on the same rough topic are likely to be pointed to by all or most of the hubs which specialise in the area. Hence even if the various authorities don't point to each other (perhaps because of commercial rivalries), they are all still linked in a fairly tight subnetwork by the hubs. Such structures can be seen as defining a *de facto* subject or topic, as created by an actual community of page authors. Such topics and communities are an alternative way of carving up the content of the Web along the lines of standard classificatory discourse [137].

4.1.2 Graph-theoretic investigations

Perhaps the best-known paradigm for studying the Web is graph theory. The Web can be seen as a graph whose nodes are pages and whose (directed) edges are links. Because very few weblinks are random, it is clear that the edges of the graph encode much structure that is seen by designers and authors of content as important. Strongly connected parts of the webgraph correspond to what are called *cybercommunities* and early investigations, for example by Kumar et al, led to the discovery and mapping of hundreds and thousands of such communities [175]. However, the identification of cybercommunities by knowledge mapping is still something of an art, and can be controversial – approaches often produce "communities" with unexpected or missing members, and different approaches often carve up the space differently [137].

The connectivity of the webgraph has been analysed in detail, using such structural indicators as how nodes are connected. Various macroscopic structures have been discerned and measured; for example one crawl of in excess of 200 million pages discovered that 90% of the Web was actually connected, if links were taken as non-directional, and that 56m of these pages were very strongly connected [49] cf. [80]. The structure thus uncovered is often referred to as a *bowtie* shape, as shown in Figure 4.1. The 'knot' of the tie is a strongly connected cluster (SCC) of the webgraph in which there is a path between each pair of nodes. The SCC is flanked by two sets of clusters, those which link into the SCC but from which there is no link back (marked as IN in the figure), and those which are linked to from the SCC but do not link back (OUT). The relationship between the SCC, IN and OUT gives the bowtie shape. The implications of these topological discoveries still need to be understood. Although some have suggested alterations to the PageRank algorithm to take advantage of the underlying topology [18], there is still plenty of work to do to exploit the structure discerned.

Indeed, the bowtie structure is prevalent at a variety of scales. Dill at al have discovered that smaller subsets of the Web also have a bowtie shape, a hint that the Web has interesting fractal properties – i.e. that each thematically-unified region displays (many of) the same

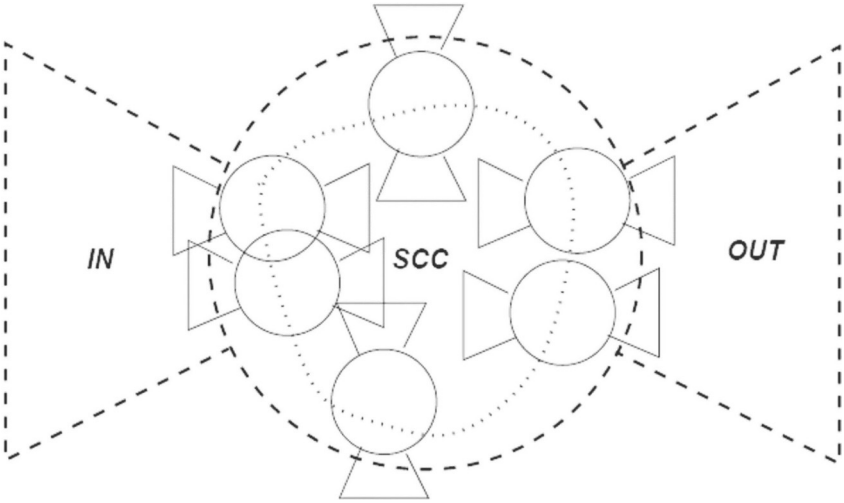

Fig. 4.1 The bowtie shape of the Web and its fractal nature [78].

characteristics as the Web at large [78]. The Web is sufficiently sparsely connected to mean that the subgraph induced by a random set of nodes will be almost empty, but if we look for non-random clusters (thematically-unified clusters or TUCs) which are much more connected, then we see the bowtie shape appearing again. Each TUC will have its own SCC, and its own IN and OUT flank, contained within the wider SCC. The larger-scale SCC, because it is strongly connected, can then act as a navigational backbone between TUCs.

In this way the fractal nature of the Web gives us an indication of how well it is carrying the compromise between stability and diversity; a reasonably constant number of connections at various levels of scale means more effective communication [29]. Too many connections produce a high overhead for communication, while too few mean that essential communications may fail to happen. The assumption that levels of connectivity are reasonably constant at each level of scale is of importance for planning long-range and short-range bandwidth capacity, for example. The Web develops as a result of a number of essentially independent stochastic processes that evolve at various scales, which is why structural properties remain constant as we change scale. If we

assume that the Web has this sort of fractal property, then for designing efficient algorithms for data services on the Web at various scales it is sufficient to understand the structure that emerges from one simple stochastic process [78].

There are a number of metrics available to graph theorists ([40] and see [76] for a recent survey). Centrality measures tell us how connected a node is compared to other nodes of the graph, and therefore can help tell us which are the most "central" nodes. The sum of distances to the other nodes (the *out distance*) and the sum of distances from the other nodes (the *in distance*), normalised for the size of the graph itself, can be informative. A central node will be one which has a relatively low total of in and out distances; in contrast nodes buried far away from central nodes are less likely to be reached by a chain of links. Knowing which are the central nodes, in particular which nodes are relatively *out-central* (i.e. there are many links from those nodes to other nodes), is an important first step to navigating through hyperspace. Such central nodes are useful for reaching arbitrary points in the graph [76].

Global metrics look at extracting information about the graph as a whole. *Compactness* is a measure of how connected the graph is; a compact graph means that, in general, it is easy to reach a randomly-chosen node from another. The usual measure has a range between 0 (totally disconnected nodes) and 1 (universal connections). Compactness of 0 is obviously hopeless for an information space, but perhaps less obviously the graph shouldn't be too compact either; if authors of webpages are sparing and thoughtful about what they link to, their links are likelier to be useful. There are also methods for discovering whether a graph is balanced or unbalanced, i.e. some parts of the graph are less well-connected compared to others, and therefore perhaps missing information. Balance is a property of an individual node on the graph, and is meant to express the intuition that, in a reasonably expressive Web resource, links can be interpreted as further developments of ideas in the resource, and that therefore if some of the links are very well connected and others rather sparsely connected, then it might be the case that the former denote a very well-developed topic while the latter could be improved with the addition of further links [40].

Other global metrics can measure the linearity of a graph, the distribution of links, or the diameter (i.e. the maximum distance between nodes). The diameter of the webgraph has been estimated at 500, and the diameter of the central highly connected core at 28 [49]. In 1999 it was estimated that the *average* distance between two randomly chosen documents was about 19 [13], increasing to 21 a year or two later [21]. The Web's structure is hypothesised to be a *small world graph*, in which the shortest paths between nodes are smaller than one might expect for a graph of that size [284].

Where a specific topic area is understood, analyses can be based on keywords, crawling the Web with a variety of search engines to produce a vicinity graph showing links between sites containing the keywords. Such graphs have been used to map scientific expertise in a number of topic areas; for instance [252] investigated vicinity graphs about climate change to determine their structural properties such as connectivity and centrality. In tandem with expert interviews, the analyses were used to uncover patterns of usage, and throw light on the question of whether the Web structure creates a democratic, decentralised science where many different suppliers of information are used, or alternatively a winner-take-all Web where existing important centres of information supply get reinforced. Their preliminary results provided some support for both of these patterns, as well as pointing up the need for data covering longer periods of time and the triangulation of expert group interviews, webmetric analysis and more in-depth case studies.

The structure and evolution of large networks have often been modelled as so-called "random graphs", whose N nodes each has a probability p of being connected to another node. The probability that a node has k links therefore follows a Poisson distribution [89]. However, in the case of the Web, it is surely unlikely that links between nodes are truly random. So, for instance, all things being equal a node will be linked to a lot of other nodes if it is well-integrated into a domain's discourse, and the challenge to graph theory is to uncover this non-random aspect of Web topology, and represent it. [21] suggests statistical mechanics as a potential source of inspiration, as it can be used to infer properties of the Web as a whole from a finite sample (even Google's index of billions of Web pages is a limited proportion).

A number of parallel studies round about the turn of the century showed that the probability of a page having k links does not, as random graph theory predicts, follow a binomial distribution and converge to Poisson for large networks; rather it decays via a power law. According to Barabási, the probability of a randomly-selected webpage having k links is k^{-G} where $G = 2.45$ for outgoing links and $G = 2.1$ for incoming links. The topological differences that follow are significant; for instance, with a network with a Poisson distribution, it will be exponentially rare to find nodes with substantially more links than the mean, whereas the power law distribution determines a topology where many nodes have few links, and a small but significant number have very many.

In the usual type of random graph, the average number of links per node is extremely important for determining structure, because of the Poisson distribution of the numbers of links. But for the type described by Barabási et al, that average is of little significance to the network; for that reason they refer to them as *scale-free* networks [22]. Barabási et al originally expected to find a random spread of connections, on the ground that people follow their unique and diverse interests when they link to documents, and given the large number of documents the resulting graph of connections should appear fairly random. In fact, the Web's connectivity is not like that. What we see is that most nodes connect to a handful of other nodes, but some nodes (hubs) have a huge number of connections, sometimes in the millions. There seems to be no limit to the number of connections that a hub has, and no node is typical of the others, and so in this sense the network is scale-free. Scale-free networks have some predictable properties, though – they resist accidental failure, but are vulnerable to coordinated attack on the hubs.

Interestingly the physical network itself is also a scale-free network that follows a power law distribution with an exponent $G = 2.5$ for the router network and $G = 2.2$ for the domain map [92]. Furthermore, it has also been reported that the probability of finding a website made up of n webpages is again distributed according to a power law [150]. The scale-free nature of the Web has yet to be properly exploited to improve significance algorithms such as PageRank. This is likely to be a potentially very fruitful area for future research [178].

The connectivity of the Web is also distorted by clustering; the probability of two neighbours of a given node being also connected is much higher than random (cf. e.g. [4]). This clustering contributes to the value of the Web as an information space, in that even random exploration of a tightly-connected cluster is likely (a) to keep the user within the cluster of relevant webpages, and (b) deliver some new knowledge or interesting slant on the topic at hand. Different types of cluster, or patterns of interaction, may produce interestingly different subgraphs with potentially different distributions. For instance, certain parts of the Web are aimed at collaborative work, such as academic disciplines (cf. [252]). Others are primarily in publish mode, as with the major media outlets. Still others are intended for personal interaction that could be quite dynamic and complex, such as a blogging topic (cf. [3, 5]). Certain suburbs of the Web will have dramatically different dynamic patterns of connectivity from each other, and from the Web as a whole.

Mapping the invariants not only brings us closer to a clear description of Web phenomena, but also enables standards for the next generation(s) of the Web to be developed that preserve the essential aspects of Web structures while allowing for growth and increases in usability, expressivity and other desiderata. For instance, understanding the network properties of the Web will help provide models for its security requirements and vulnerabilities, its tendency for congestion, the level of democratisation it will support, or what would happen if a 'two-speed' Web came into being as a result of preferential treatment being offered to certain Web users and the ending of net neutrality. See Section 4.2.4 for further discussion of the practical application of mapping the Web.

Traditional graph theory tends to work with models of a fixed size. However, the growth of the Web not only demands a dynamic graph theory, it also requires models that respect the quality of that growth. So, for example, new links are not randomly distributed, any more than the old links were; the probability is that a new link will be connected to pages that are themselves highly connected already (thus displaying *preferential connectivity*). Given that constraint, Barabási et al have modelled Web-like networks as graphs in which a new node gets

added at each time step, whose links to other nodes are distributed non-randomly, with a greater probability of connection to highly-connected nodes. Such a graph is also scale-free, and the probability that a node has k links is a power law once more, with exponent $G = 3$. In such models, highly-connected nodes obviously increase connectivity faster than other nodes [21].

Such scale-free models are simple examples of evolving networks – are they too simple? In particular, the power law assumption [92] might be too neat, and the distribution of node degree, though highly variable, may not fit a power law [59]. Alternative models are beginning to emerge [94]. One important line in Web Science should be the exploration of dynamic graph topologies, to investigate how the peculiar patterns of Web growth could happen, and how they might be modelled. Furthermore, the effects of scale are still not understood. Is there some kind of upper limit to the scalability of the Web? If so, is that limit a principled one, or does it depend on the availability of feasible technology? How large can the Web grow while remaining a small world in the sense described above.

Indeed, questions of scale cut both ways. There are other, smaller Webs around, and whereas the Web itself came as something of a surprise to mathematicians and computer scientists when it began, now Web studies tend to look mainly at *the* Web. Structures such as Intranets have very different properties, in terms of size, connectivity, coherence and search properties; some properties carry over from the Internet as a whole, while others don't. There has been little work on these contrasting structures, though see [91] for investigation of Intranets, and [252] for the subgraphs corresponding to particular scientific topics.

4.2 Web mathematics

López-Ortiz, in a useful survey [191], looks at a number of paradigms useful for understanding the *algorithmic* foundations of the Internet in general and the Web in particular. Applying insights about algorithms to networking problems, in the context of the specific protocols underlying the Web, is potentially very fruitful. And that context is

vital – the functioning (or otherwise) of algorithms in the context of the Web provides some of the most convincing evidence for those who wish to argue that it is an importantly unique environment. The growth of the Web, as López-Ortiz points out, was such that the most advanced text indexing algorithms were operating well within their comfort zones in standard applications at the beginning of 1995, but struggling hard by the end of that year.

4.2.1 Rational models

One important paradigm is that of microeconomics, discrete mathematics, rational choice theory and game theory. Though individual users may or may not be "rational", it has long been noted that *en masse* people behave as utility maximisers. In that case, understanding the incentives that are available to Web users should provide methods for generating models of behaviour, and hence insights into what global sets of desirable behaviour can be engineered, and what systems could support such behaviour.

The Web has no central coordination mechanism, yet produces systematically interesting behaviour thanks to incentives and constraints imposed either by architecture, protocols and standards, and their interaction with social or psychological properties of users or designers (indeed, it is arguably the fact that the Web is built, run and used by a multitude of real-world users with almost unimaginably diverse interests and preferences that is of greatest significance for the application of the economic/game theoretic paradigm). Are there upper limits to the utility of the freedom that decentralisation has produced? As the number of users increases, will the chances that the choices that one makes impinge on the range of choices available to others increase, or is that an illegitimate extrapolation from the real world with fixed spatial parameters? The answer to that question, however mathematical, will have profound effects on Web governance [186]. Put another way, what is the frequency with which Nash equilibria are discovered which are suboptimal for all parties? In a decentralised and growing Web, where there are no "owners" as such, can we be sure that decisions that make sense for an individual do not damage the interests of users as a whole?

Such a situation, known as the 'tragedy of the commons', happens in many social systems that eschew property rights and centralised institutions once the number of users becomes too large to coordinate using peer pressure and moral principles.

The key to the success of the Web lies in the network effects of linking to resources; if a good has a network effect, then the value of that good increases to its individual owners the more owners there are, and all things being equal the richer the set of links the more use linking is. Network effects can either be direct or indirect. A direct effect is where demand for a good is connected to the number of people who have it – telephones and emails being prime examples. Intuitively, we can see that modelling markets for such goods is problematic, as demand seems to depend on a number of apparently unrelated decisions (to adopt or not in the early stages); if 'enough' people go for it early on the market will soar, otherwise not. But how do we define 'enough' here? Put more technically, what this means is that the market with network effects has multiple equilibria. As the number of adopters (size of the network) increases, the marginal willingness of consumers to pay increases because of the greater gains they will receive from the service for a given price – such gains, dictated by the actions of third parties rather than the two parties to the actual transaction, are called *positive externalities*. But beyond a certain threshold, the willingness to pay falls off, as the later adopters typically get less from the network.

So, for instance, consider a subscription VOIP service with free calls to fellow subscribers. A small number of subscribers generally reduces the value of the service to a potential user, but if we assume the price stays steady, if the number of users increases, the number of people prepared to pay the price will increase, and there will be a virtuous circle of growth. However, those joining later will be those who are more sceptical about the value of the service – it may be that they don't particularly have much need for VOIP. So at some point a maximum will be reached, when even a very large network, with a lot of communication possibilities, will not attract any new users without a lowering of the price. Many online services have this network structure, for instance mobile networks or interactive poker or gambling sites.

If, as in Figure 4.2, the supply curve is perfectly elastic (i.e. horizontal), there are three equilibria: the two points where the supply curve crosses the demand curve (at network sizes B and C), and the point at which the supply curve hits the y axis ($A = 0$). If the network size stays at 0, then demand remains nil, and we stay in position A. At C, the position is also stable; the network contains all the customers prepared to pay the market rate, and cannot grow as there is no-one else prepared to pay. If the network grows, it must be because the price has fallen (i.e. the supply curve has moved downwards; if the network shrinks, that must be because someone has changed their preferences and is now no longer prepared to pay the market rate (i.e. the demand curve has moved downwards). If we assume that the two curves remain stationary, then any change will result in a slip back to C. The key point is B, which though an equilibrium is unstable. If the network size slips below B, then not enough people will be prepared to pay the market rate and the demand will gradually slip back to zero. If on the other hand it can get beyond B, then suddenly many more consumers will appear who are prepared to pay the market rate or more, and the network size will increase dramatically, getting over the demand curve's hump and reaching C. Hence B is a critical mass for the network [281].

Interpreting this graph in Web terms, 'network size' could be glossed as 'number of nodes in the webgraph' or alternatively 'number of links'. 'Willingness to pay' refers to the costs that the Web user is prepared to absorb. These include regular financial costs such as the hire of a broadband line, upfront financial costs such as the purchase of a computer, upfront non-financial costs, such as the effort involved in ascending learning curves associated with particular formalisms or applications, and regular non-financial costs such as constantly ensuring that one's system is secure. The 'users' being referred to will also vary: the graph could refer to ordinary web users (consumers of content, whose costs will typically be financial), but might also refer to web authors (creators of content, whose costs will typically be in terms of time and effort). But either way, the continuation of the positive network effects observable on the Web depends upon sustaining performance beyond the second, unstable equilibrium.

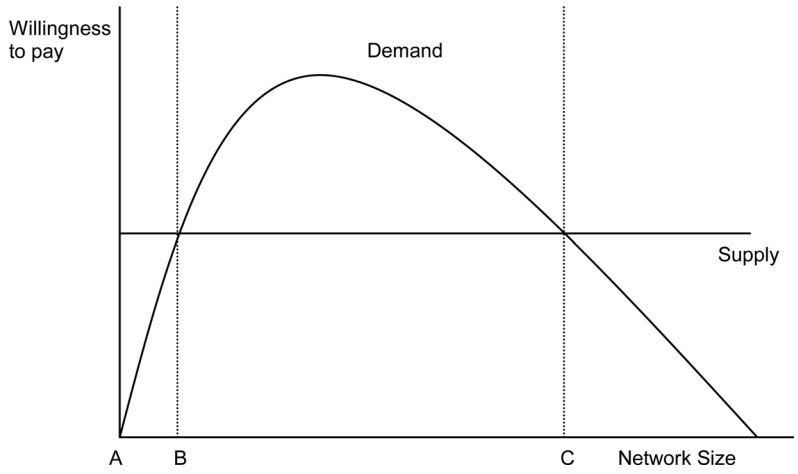

Fig. 4.2 Demand and supply for a network good [281].

Indirect network effects also apply to the Web. An indirect network effect is found in an industry such as DVDs – my purchase of a DVD player is unaffected by who else has one, but the larger the number of DVD player owners, all things being equal the larger and richer the amount of DVD content available will be (and indeed the cheaper it will be). Modelling such indirect effects is also an important part of understanding how the Web can continue to grow.

How easy will it be to describe the Web in game theoretic/rational choice terms? Are there intrinsic differences between, say, 'ordinary' users and service providers? And again, how do we understand, on this paradigm, the growth of the Web and the invariants of the Web experience? This is key to modelling the evolution of players' views given the feedback they receive from experience. How do we assess the fixed points in the system? Or construct equilibria for particular game setups? Or design mechanisms to enforce "good" behaviour? Or model the evolutionary behaviour of groups in such large scale systems? Perhaps most importantly, how do we undertake the inverse game theoretical problem of identifying equilibria of prohibitive cost and engineer mechanisms to prevent them coming about?

Answers to such questions appear on (at least) two levels. First of all, the behaviour of users in terms of (neutrally-conceived) demands for information needs to be coordinated within the capabilities of the physical network of information flow along actual physical wires. Coordination and routing of information needs to happen without friction, and game theory should be of value in modelling that. And secondly, the interpreted behaviour of Web users needs to be such that the potential for deceit and other costly forms of behaviour is minimised. There is no engineering solution to the problem of trust (see Section 5.4.4), but on the other hand there may be ways of engineering the Web so that trustworthy behaviour can be justly rewarded without imposing too many costs on users or reducing the number of interactions so drastically that the beneficial network effects are minimised.

4.2.2 Information retrieval models

A second important paradigm is that of information retrieval. IR is the focus for an arms race between algorithms to extract information from repositories as those repositories get larger and more complex, and users' demands get harder to satisfy (either in terms of response time or complexity of query).

One obvious issue with respect to IR over the Web is that the Web has no QA authority. Anyone with an ISP account can place a page on the Web, and as is well known the Web has been the site of a proliferation of conspiracy theories, urban legends, trivia and fantasy, as well as suffering from all the symptoms of unmanaged information such as out-of-date pages and duplicates, all the difficulties pertaining to multimedia representations, and all the indeterminacies introduced by the lack of strictly constrained knowledge representation. Understanding exactly what information is available on a page waiting to be retrieved remains a serious problem.

Perhaps more to the point, traditional IR has been used in benign environments where a mass of data was mined for nuggets of sense; typical problems were complexity and lack of pattern. Benchmark collections of documents for IR researchers tend to be high-quality and almost never intentionally misleading, such as collections of scientific

papers in particular journals. Other Web-like mini-structures that can be used, such as Intranets, are also characterised by the good faith with which information is presented. But malicious attempts to subvert the very IR systems that support the Web so well are increasingly common. Web-based IR has to cope with not only the scale and complexity of the information, but potential attempts to skew its results with content intended to mislead [139].

4.2.3 Structure-based search

The IR result that really brought search into the Web age was the discovery that it was possible to make a heuristic distinction between those links that appear to denote quality of the linked-to site, and those that do not [171, 221], based only on the computation of the eigenvalues of matrices related to the link structures of local subgraphs. Neither Kleinberg's HITS algorithm nor Page et al's PageRank requires any input other than the otherwise uninterpreted structure of hyperlinks to and from webpages.

The duplication problem is interesting in the context of this paradigm. What methods can be found for identifying duplicate pages when hyperlink structures may have changed dramatically, and when other aspects of content such as headers, footers or formatting may have changed as well [76]? Could such methods be helpful in uncovering cached pages that are otherwise unavailable in their original location? Would the resulting persistence of information in webpages actually be a good thing, given that the maintenance of online information repositories is already one of the major costs of Web-based knowledge management? Evaluating the effectiveness of Web search and retrieval techniques, particularly given the money to be made from search [25] – Google's IPO in 2004 valued the company at around $30bn in a faltering stock market – is naturally the focus of much research. Metrics for engine performance are appearing all the time, focusing on the effectiveness of the search, and the comparison of different engines [76].

The aim of search is to retrieve pages that are relevant to the user's query, i.e. those pages which, when accessed, either provide the reader with pertinent information, or point the reader to other resources that

contain it. So one can look at IR-based measures for a search engine's *precision* – in other words, the proportion of returned pages that are relevant – or *recall*, the proportion of relevant pages that are returned (cf. [280]). It goes without saying that what search engines themselves search for (at the metalevel, so to speak) is the magic combination of high precision and high recall – although determining recall involves determining, at least approximately, the number of relevant pages across the Web as a whole, which is needless to say a particularly hard problem.

Search engines must also struggle to remain current, by reindexing as often as possible, consistent with keeping costs down, as the Web grows and individual pages are edited or changed as the databases underlying them change [43]. Search engines can be compared using various parameters, be it their coverage (the number of hits returned given a query, particularly looking at the number of hits only achieved by that search engine); the relevance of the pages returned; the time taken; or the quality of returns. As one would expect, different engines do well on different metrics [76].

4.2.4 Mathematical methods for describing structure

Understanding the mathematics and topology of the Web is of practical import for understanding the invariants of the Web experience and therefore providing roadmaps for extensions to the Web. One important property that the Web possesses is robustness in the face of undermining influences; neither hackers nor the inevitable faults in the physical network greatly disrupt the Web, even though something like one router in forty is down at any one moment. Barabási and colleagues [253] advocate the use of *percolation theory,* the study of processes in idealised random 2 (or more) dimensional media [119], to look at the topological contribution to fault tolerance. For example it has been shown that for scale-free networks, for a connectivity exponent $G < 3$ (on the assumption of node connectivity being distributed according to a power law), randomly removing nodes will not fragment the network into disconnected islands [63]. As we have seen, on the assumption that the Web is a scale-free network with power law distribution, the exponent G is

significantly less than three, and so the Web should be very hard to fragment (though [63] focused on showing the resilience of the Internet as a whole). The theoretical results back up empirical computer simulations that show that removing up to 80% of the nodes of a large scale-free network still leaves a compact connected cluster [21].

On the other hand, percolation theory shows that scale-free networks are somewhat more vulnerable to directed, coordinated attack, even if they are robust against random failure. Non-random failures could be damaging if they targeted the highly-connected sites in particular; failure of a small number of hubs could dramatically increase the diameter of the Web (in terms of the smallest number of clicks needed to go from one randomly-chosen page to another), and failure of a significant number of highly-connected sites could lead to fragmentation [64].

4.2.5 Mathematical methods for describing services

As the Web evolves to include a services model, where software agents and Web services will live online and be invoked by users, and where an increasingly important metaphor is that of the client contacting a service provider, new mathematical representations, formalisms and theories are becoming useful to describe this relationship.

The theory of *Petri nets* [269, 298] models discrete distributed systems, of which the Web is a prime example. The theory in effect adds the notion of concurrency to the idea of the state machine, and has been suggested as an important means of modelling Web services [296]. *Process algebras,* such as CSP [141] or CCS [203] can also model parallel processing. They provide an array of constructs to model the dynamic processing of information and communication of outputs and requested inputs, such as actions, sequences of actions, choice functions, processes and methods of synchronisation.

One recent development is the π-*calculus* (named analogously to the λ-calculus), which is a development of process algebra (specifically an offshoot of CCS) designed to provide mobility in the modelling of processes. The π-calculus is intentionally minimal (containing little more than communication channels, variables, replication and concurrency),

but can be extended easily to encompass first order functions and basic programming constructs [204, 1].

As we have seen (Section 3.3.1) there is a need for languages to describe web services (such as CDL or BPEL), and it may be that the mathematics listed here could underpin such languages. There is a lively debate about Petri nets and the π-calculus [24], focusing on the relative merits of the graphical, state-based nets, and the more textual, linear, event-driven process algebras [276].

5

Social Aspects

The Web is a piece of computing embedded in a social setting, and its development is as much about getting the embedding right as it is doing the engineering. In this section we will look at the social, cognitive and moral context of the Web, and discuss ways in which social requirements can feed into engineering decisions. This discussion does not include the enforcement of standards or institutions of governance, which are covered in Section 6.

5.1 Meaning, supervenience and symbol grounding

The Web is often conceived as a set of layers, with standards, languages or protocols acting as platforms upon which new, richer, more expressive formalisms can sit. Such platforms, like TCP/IP, are deliberately intended to be as neutral as possible. The Semantic Web is an obvious example of a layered yet unprescriptive architecture [32].

Such layered representations are not *reductive* – that is, the upper levels are not merely shorthand for expressions at the lower levels. But there is an interesting question to do with the significance of such layered representations of the architecture. In particular, the nearer to the

top that an expression is found, the more likely it is to have *meaning*. By which we mean that although an expressive language needs to have formal syntax (and possibly semantics), to be significant it still needs to map onto human discourse in an intelligible way.

In the Semantic Web model, ontologies are intended to perform this mapping, and to make meaningful dialogue between human and machine possible [97], although it is important to be clear that such mappings aren't *magic*: ontologies, as artificial creations, stand in just as much need of mapping onto human discourse as the structures they map [113, 289]). And in this, they are no different to other structured formalisms, such as queries [39].

One view is reminiscent of the philosophical idea of *supervenience* [168, 169]). One discourse or set of expressions A supervenes on another set B when a change in A entails a change in B but not *vice versa*. So, on a supervenience theory of the mind/brain, any change in mental state entails *some* change in brain state, but a change in brain state need not necessarily result in a change in mental state. Supervenience is a less strong concept than reduction (a reductionist theory of the mind/brain would mean one could *deduce* mental state from brain state, that psychology follows from neuroscience). And it has been thought over the years that supervenience is a good way of explaining the generation of meaning: uninterpreted material in the lower layers of discourse is organised in significant ways so that the material in the upper layers is constrained to be meaningful. It may be appropriate to think of the Web as having this sort of supervenience layering: the meaningful constructs at the top depending crucially on meaningless constructs in HTML or XML or whatever below.

If we are to see the higher levels of the Web as supervenient on the lower, then the question arises as to what the foundational levels of the Web *are*, and the further question of whether they *have* to take some particular form or other. One does not have to subscribe to the requirement for *symbol grounding* (i.e. the need to avoid symbol meaning being 'grounded' only in other symbols, and instead being grounded by some direct relationship with a referent – [129, 130] – a requirement that Wittgenstein, among others, denied could be fulfilled – [291]) to

expect to see some sort of discourse of uninterpreted symbols playing a foundational role.

'Meaning is use' is a well-known slogan that represents a key insight in Wittgenstein's later philosophy of language. It clearly contains a very important insight, and applied to natural language is a powerful message to understand meaning in terms of what people use their language to do. The same insight applies to the Semantic Web, but there is a wider question of what 'use' consists in. In the world of machine processing and interoperability of data, much of the use or discourse is automatically generated by computers. For that reason, it is not clear that definitions in words, or code, or quite specific uses, will not suffice to pin down terms for the Semantic Web with sufficient accuracy to allow logical deduction to take place. Stability of the referents of key URIs, for example, might enable a great deal of automation in specific topic areas – a notion of science as underpinning meanings reminiscent of the theories of Hilary Putnam [233]. The fact that the Semantic Web works in the world of relational data, with machines doing much of the work, means that it isn't necessarily incumbent upon it to solve the problems of definition and logic that have proved so resistant to analysis in the world of natural language, although new insights may be gained from the grounding in URIs discussed in section 3.1.2 above.

5.2 Web reasoning

5.2.1 Plus ça change?

As we have seen, there are various issues in the science of the Web with semantic, philosophical or logical roots. This is not the first time that practitioners of a computational paradigm have suddenly had to familiarise themselves with Philosophical Logic. The general project in Artificial Intelligence (AI) of trying to produce general adaptable problem-solvers on the basis of symbolic descriptions and reasoning, a strong (and *prima facie* reasonable) driver of AI research through the 1960s and 1970s, ultimately foundered on the difficulties of specifying everything required for computers to reason about arbitrary situations. This failure led to the disparaging name 'GOFAI' (Good Old Fashioned AI) for the project.

Some argue that GOFAI is hindered by the failure to solve the frame problem, the fact that real-world reasoning seems to be highly situated, and that any description or representation can never be restricted to terms with local significance – to understand anything a computer would have to understand everything [133, 82]). Others say that AI cannot reason *about* anything until there is a solid connection between the terms with which a computer reasons and its referents, a connection not provided by the programmers' programming it in [129, 130]. There have also been claims about the type of thing a computer or robot is, although criticisms of the hardware are less important here than the shortcomings of the semantic and logical underpinnings of GOFAI (cf. [255]).

There are, it has to be said, AI-independent arguments that would seem to support the GOFAI project, that (for instance) 'knowing how' is merely a species of 'knowing that', and that procedural knowledge is, whatever the appearances, a relation between an agent and a proposition [267], but such arguments do not seem to be borne out by the technology. An alternative to GOFAI, it is argued, are relatively dumb methods based on syntax and numerical computation – these 'unintelligent' methods (such as PageRank, IR, NLP) turn out to behave much more effectively.

It is argued by some that the Web – and specifically the Semantic Web project – threatens to make all the same mistakes as GOFAI. In particular, the need to create ontologies to aid data sharing and so on has been seen as requiring a context-free theory of everything [158]. The widely cited CYC project, to produce a giant knowledge base and inference engine to support 'common sense' reasoning [183] does not seem to have broken the back of the problem, while the ontologies produced by the formal philosophical ontology movement [124, 106, 210] seem somewhat complex and daunting, although it has been suggested that they may be used (as a sort of 'deep' ontology) to bring together overlapping lightweight ontologies and relate them to each other [228]. Ultimately, goes the argument, it is the situated nature of human cognition that makes it possible for the human mind to do exquisite processing of distributed and multimodal knowledge.

On the other hand, the claim that the Web, and the SW in particular, will hit the same problems as GOFAI needs to be seen in the context of the manipulation, sharing and interrogation of relational data as envisaged by the SW programme. Data are already shared and amalgamated in a number of contexts by special purpose applications, which stitch together underlying ontologies with relevant mappings and translations. These translations need not be universal, and need not aim to produce a globally consistent ontology. The SW generalises this sort of approach to shared data systems by developing standards for mapping between data sets; further argument is needed to establish that this programme will fall foul of the standard objections and practical obstacles to GOFAI.

In particular, the SW does not rely on, nor necessarily aspire to, the production of the level of intelligence envisaged by GOFAI theorists. Partial solutions will work, and will be aimed at, on the SW. It would of course be good if an artificial agent could produce the range of inference that a human might, but that is not an explicit goal of the SW, and the SW will not fail if such an agent is not produced. The aim is to produce an extension to the Web that will enable more information to be produced more easily in response to queries. GOFAI was aimed at producing an intelligent system exhibiting human-level intelligence; the SW should assist something of human-level intelligence (usually a human) in everyday information discovery, acquisition and processing [17].

There have also been arguments that ontologies seem less problematic when viewed from this perspective. Bouquet et al describe C-OWL (or Context-OWL), an extension of OWL that allows context-dependent ontologies to be represented [41]. And at least one commentator has seen the SW as a potential saviour of the expert system research programme. Whereas large knowledge bases force knowledge into a straitjacket, where things that don't fit don't get represented, and knowledge representation is a piecemeal affair driven by the contingencies of the KRL, the SW provides the means to much greater flexibility of representation and capture of rationale. Resilient hyper-knowledge bases, containing many links out and multiple representations of the

same or related knowledge should be more adaptable to change and reuse [110].

5.2.2 Alternative ways of reasoning

There are many different types of reasoning, but not too many have been successfully automated beyond deductive linear reasoning and various statistical methods. What alternative methods has the Web facilitated? One obvious candidate is *associative* reasoning, where reasoning on the basis of associations – which can be extremely unpredictable and personalized – takes one down a train of thought [202]. So, for example, the classic case of associative reasoning is given in Proust's novel *Remembrance of Things Past*, where the middle-aged narrator, upon eating a Madeleine dipped in tea, finds himself transported to his childhood in Combray, when his Aunt Léonie would give him a Madeleine on Sunday mornings. On the Web, the potential of associative reasoning is immense, given the vast number of associative hyperlinks, and the small world properties of the Web. Google-like searches, valuable though they undoubtedly are, cannot be the whole story in a world of small pervasive devices, software agents and distributed systems [127].

However, associative reasoning via hyperlinks, though an attractive and important method, is not the only way to go about it. This type of reasoning is not strictly associative reasoning proper, as the associations are those of the *author*, the person who puts the hyperlinks into a document. In Proust's scene, this is like Marcel taking a bite of his Madeleine and suddenly and unexpectedly perceiving the memories of the *baker*. Open hyperlinking allows the *reader* to place link structures over existing Web pages, using such information as metadata about the page in question, relevant ontologies and user models [54]. Associativity is clearly one of the major driving forces of the Web as a store of knowledge and a source of information. Associative reasoning, for example, has been used for collaborative filtering in recommender systems [177].

Another type of reasoning is *analogical* reasoning, another highly uncertain type of reasoning that humans are remarkably successful at using. Reasoning by analogy works by spotting similar characteristics between two subjects, and then assuming that those subjects have more

characteristics in common – specifically that if subject A has property P, then by analogy so does subject B [109]. Obviously the success of analogical reasoning depends on having representations of the two subjects which make it possible to spot the analogies, and in being suitably cautious (yet creative) in actually reasoning. Case-based reasoning (CBR) is a well-explored type of analogical reasoning.

Analogical reasoning can be made to work in interesting contexts [199], and reasoning engines exist [266]. Sketches of an approach using analogical reasoning to generate metadata about resources have appeared recently [299], and case-based explanations can be useful in domains where causal models are weak [214]. In a domain described by multiple ontologies, analogical reasoning techniques may well be useful as the reasoning moves from one set of ontological descriptions to another, although equally the change of viewpoint may also complicate matters. There have been interesting attempts to support analogical reasoning (i.e. CBR) across such complex decentralised knowledge structures [70], and also extensions to XML to express case-based knowledge [66].

5.2.3 Reasoning under inconsistency

The Web is a democratic medium. Publishing is cheap, but that means that we should expect inconsistency. For the Web the classical principle of *ex falso quodlibet*, that the conjunction of a statement and its negation entails any proposition whatever, is clearly too strong. Enforcing consistency checking and trying to outlaw contradiction is a non-starter thanks to the social pressures towards inconsistency on the Web, or indeed other large-scale distributed systems. The likelihood of errors (incorrect data entries) is of course high. Malicious or mendacious content will exist. But most importantly, there will be serious disagreements in good faith in all sorts of areas. These social forces make inconsistency inevitable across any decent-sized portion of the Web – and indeed have already driven a great deal of reasoning strategies in AI, where systems were designed in the expectation of having to cope with contradictory knowledge bases, or where the possibility exists that a statement that was true in a model at one point might not

be true further on. Such strategies assume that inference is situated, and that the desirability of discovering and exposing contradictions is dependent on context (cf. e.g. [140] for an early example from AI).

The major advantage of classical logic is that it scales. Hence one solution to the inconsistency problem is to evolve strategies for dealing with contradictions as they appear. For instance, something on the Web is asserted by some formula in a document, but different documents need not be trusted to the same extent. Associated with documents will be metadata of various kinds, which may help decide whether the statement in one document should override its negation elsewhere.

Alternatively, this is an application opportunity for *paraconsistent logics*, which allow the expression of inconsistencies without the corresponding deductive free-for-all. Paraconsistent logics localise the effects of inconsistencies, and often require semantic *relevance* of propositions used in deductions (the proof of *ex falso quodlibet* requires the conjunction of an irrelevant proposition with the contradictory ones), which prevents the effects from spreading beyond the contradictory hotspot [15, 262], and see [231] for a survey).

Other approaches include having multiple truth values to complicate the analysis of a contradiction (and the appearance of contradiction may indeed often be due to all sorts of contextual factors that are very hard to analyse and formalise). And one of the few types of paraconsistent logic with a respectable implementation history as well as clean semantics and proof theory is annotated logic [95, 271]). Modal logics, which might treat Web resources as possible worlds within which inconsistency was bad, but between which was allowed, would be another angle; certainly this approach is important within the agents community [e.g. 270].

In Web Science terms, the issue of the "correct" logics for the Web will depend on context, purpose of analysis and so on. But it is clear that modelling the Web is essential for a number of purposes where proofs are required about what is entailed by a series of statements (for example, in discovering whether information has been used correctly or incorrectly – cf. [287]). And in the SW, logic plays a larger role. Which logics are appropriate for the Web, or the SW? What problems of scale should we anticipate? Are there *ad hoc* methods that might get round

the deep logical issues, and allow relatively straightforward logics to function with a restricted zone of application? And how can standards work, and the setting of standards help resolve logical issues?

5.3 Web epistemology

Computers have revolutionised epistemology, and the Web most of all. Ideas such as the Semantic Web hold out the possibility of an extension of automation of information processing. The e-science movement has proved especially interesting. Philosophically, scientific method has proved hard to nail down, but this was partly because the logical structure of research and inference was inevitably undermined by the human and collective nature of the process, which entailed that social processes, political processes and discovery heuristics were at least as important as the logic.

Furthermore, by allowing annotation about provenance and other underlying knowledge generation issues, the Web allows a strong and institutionalised appreciation of the context of knowledge (what it assumes, what methods created it, and ultimately what political and social ends the knowledge was developed to serve). Such metadata are often important in the heuristic evaluation of knowledge, and the Web provides an opportunity to understand the history of a piece of knowledge, and the contribution that that history makes to its trustworthiness [110].

There are two important epistemological questions for Web Science. The first is what properties will future platforms need to have in order to allow as much information as possible to gravitate to the Web without imposing structure or governing theories upon it? One aim of the Web is to facilitate rational discussion of ideas, rather than the sorts of rancorous *ad hominem* attacks that make up rather too much of what is loosely called debate [30].

And secondly, the Web has a radically decentralised structure. Given that, of course it can be used frivolously or maliciously. How can we make it more likely rather than less that good science and good epistemology ends up in the Web, and not superstition? Indeed, is that a good thing? By and large, most people behave in good faith with

respect to each other in most walks of life. And opinions differ, even in good faith. But there is a constant trickle of evidence that the Web is being used to cement opinions, in polarised political situations [3], in marginalised groups [272], and even in terrorist circles [245, 285]. Can we find the best balance between free exchange of opinion and restricting opportunities for deliberate self-marginalisation?

5.4 Web sociology

The Web is a mirror to human society, and reflects the interests, obsessions and imperatives of 21st century human existence extended over a very wide range (perhaps the widest range of any human information space) of value sets, cultures and assumptions. Analysis of the search terms typed into Google is likely to be a key resource for historians of the future. In this section we will look at the relationship between the Web and its users, readers and authors. What do people and communities want from the Web, and what online behaviour is required for the Web to work? These are especially difficult questions given the radical heterogeneity of the Web – some people want to use the Web for information sharing, some for leisure and entertainment, some want to exploit the distributed information on the Web to perform science in radically new ways, others want an arena for commerce, while still others wish to create and people the sort of anarchistic utopia that has proved so elusive offline (cf. [186]).

5.4.1 Communities of interest

The Web has spawned a number of interesting and novel communities with intriguing properties. For instance, Massively Multiplayer Online Role-Playing Games (MMORPGs), where a publisher provides a persistent online space in which a game takes place, have spawned giant economies and codes of virtuous conduct as very large communities of players (sometimes of the order of millions) spend increasingly large amounts of time online [55]. The potential for behaviour in such communities will of course depend to a large extent on what the architectures allow [186], and the sizes of such communities can be very large. As early as 2001, it was reported that 84% of American Internet users

(90m people) used the Internet to stay in touch with some sort of group; that report, by the Pew Internet research project, is very informative about the ways that Americans use the Web to remain in touch with all sorts of interest groups [145].

We have already looked at some structure-based methods of uncovering cybercommunities; communities can also be studied by looking at communication between the members and the knowledge they share [185]. Proxies for trust (for example, collaborative working or email networks) can also be used to map the spread of communities of interest or practice [12, 216, 151, 297], which can have real benefit in a number of areas. For example, the evaluation of funding programmes designed to foster interdisciplinary research can be supported by evidence for the formation or otherwise of a new community by looking at patterns of collaborative working [10]; directed email graphs have been used to identify leadership roles [151]; potential conflicts of interest between authors and reviewers of scientific papers have been monitored using patterns of acquaintance in social networks [14]. A study of political blogs in the US election of 2004 showed interesting patterns of behaviour characteristic of the liberal and conservative political commentators; the two sides found different news items significant, and linked much more tightly to ideologically congenial sites, although conservative bloggers linked more densely both to each other and the liberal opposition [3]. This finding is in line with the predictions of legal scholar Cass Sunstein [272] about the behaviour of people in an online world where personalisation of content is possible and routine, although a recent survey of leading experts showed that such predictions remain controversial and disputed [103].

The Web and the Internet in general have underpinned new types of interaction, and provided a 21st century perspective on some old ones. Recent surveys have discovered large increases in the numbers of people selling something online [184], using search engines [236], using webcams [237] and listening to podcasts [238]. The Web, and other new technologies such as pervasive computing have allowed new conceptions of space to develop and supported new methods of interacting online (cf. [55]), or new interactions between virtual space, physical space or theoretical or measured spaces such as maps and plans [79]. Web

interactions are important with respect to existing communities in three ways: increasing transparency, allowing offline communities to grow beyond their 'natural' boundaries, and allowing different, more codified, types of communication between community members [71].

In general, the promotion of new methods of interacting and cementing communities – and indeed new types of community – is one of the aims of the next generation Web, as part of a more general aim to increase the amount of material on the Web, to make it more relevant to more people, and to get people to add resources to the Web without being forced to. It is impossible to predict exactly what community or interaction types will develop, but all things being equal leaving as many options as possible open should facilitate novel developments (cf. [186]).

People will use new platforms in novel and unpredictable ways, which often evolve over the long term (making them hard to observe even while behaviour is changing). Furthermore, following trends first hand involves observing users in their everyday environments; lab conditions are inappropriate [108]. Hence understanding what engineering and architectural requirements are placed on the Web by the needs of communities is a challenging problem [277]. And Web Science needs not only the effective analysis of user interactions "in the wild" so to speak; this needs to go hand in hand with the development of theories (both at the sociological and technical levels) about what it is about successful participative technologies such as RSS, folksonomies, wikis and blogs, that is common across the space. And, last but not least, what interfaces are important?

5.4.2 Information structures and social structures

The social structures of the Web depend on the *engineering* structure that underlies its upper level fabric. The understanding of the relation between humankind and technology, of the implications for society of humans being tool-using animals, has been a feature of much philosophical, political and social commentary of the post-Enlightenment period, for example in the work of Marx and Heidegger. The Web is a reflection of human intellectual and social life, but is also specifically engineered to be a tool.

In particular, the Web's structure is an advance on other more traditional data structures. All large-scale data structures end up inevitably with some form of congruence with their human context, for example within an organisation or corporation [50]. Information hierarchies, for example, have developed within hierarchies as informational structures that met certain of the embedding organisation's needs. The problem with hierarchies is that information tends to be used in the context it was originally created for. Reuse of information is often problematic, but in a very hierarchical organisation with a correspondingly hierarchical information system, the problem is finessed by knowledge not typically being retrieved outside the context for which it is created (cf. e.g. [295]).

This is not to say that trees are necessarily a bad type of structure; the tree-oriented world of XML was an improvement on the line-orientation of UNIX. Trees allow many important possibilities, such as top-down structured design, information hiding, and a degree of control combined with flexibility. But behaviour within that sort of structure is constrained: GOTO statements are considered harmful, for instance, because control of processing is lost, and the analysis and verification of programs becomes arbitrarily hard (cf. [77]). To get from one part of an information hierarchy to another one typically has to ascend until a node common to each subtree is reached and then descend the second subtree. For structured, centralised environments where control is important, this is an important innovation.

The engineering innovation of the Web is what creates added value for its human users. The development of URIs allows the speedy and unconstrained traversal of the information space in any direction; from any point in webspace one can reach any other point immediately (one can choose to be constrained by following links or the output of search engines, of course). In other words, GOTO is reinstated; global GOTOs are legitimised, because when such movement is allowed the possibility is opened of *serendipitous reuse*. Reuse in predictable situations, as occurs with hierarchical information structures, can also happen on the Web, and GOTOs have their costs. The analysis of interaction and cooperation is harder, as Dijkstra predicted, and also the system depends on the upkeep and proper functioning of the URI space.

Similarly, the simplicity of providing a 404 error page when no resource exists at the URI proffered is another important engineering factor; the browser has successfully communicated with the server but the server was either unable or unwilling to return a page. The key is that the display of a 404 error is a lightweight response to a failure that doesn't hinder the user's activities in any way; pressing the 'back' button on the browser restores everything to the *status quo ante*. The Web could not function without this error tolerance.

Information structures are not the only socially-based structures on the Web; other users have a more process-oriented set of requirements. For many the important issue is not sharing information but rather sharing *know-how*; for such users, the key is not so much to provide ontologies as ways of expressing workflow. And modelling information flow rather than state has provided an interesting route into the creation and discovery of Web services [208, 98]. On the other hand, as in other areas, ontologies and workflows are not incompatible, though they address different issues; indeed, workflow development can be ontology-mediated [225]. And judicious modelling of workflow can produce distributed and dynamic task understanding (e.g. for Web service composition) that avoids over-centralisation of workflow enactment [282].

The information the Web provides is used in useful processes embedded in human societies, perhaps most obviously productive human work. Understanding the way that information and Web technologies are used for specific purposes is an important goal for Web Science. Data about this can be hard to come by, but when data sets do become available, extremely interesting research can be done (such as [52]).

5.4.3 Significance and its metrics

A related concept to the use of a particular Web resource in a process is its *significance*. One can guess the significance of a page to a user or a community to some extent intuitively: one might expect US taxpayers to be relatively more interested in an IRS FAQ page than an arbitrary page, a Goth in Nine Inch Nails' homepage, and conservative women in angryrepublicanmom.com. There are a number of methods of fixing the various potential interpretations of such intuitions via some hard

mathematics, which is a good way to begin to understand the social dimension of the Web. And understanding the significance of a page is important for the non-trivial task of ordering pages retrieved during Web search-and-retrieval.

Significance can be decomposed into two types of metric: *relevance* and *quality* [76]. Relevance is connected to the idea of querying: how many queries does a page handle? The different ways of answering that question have led to the development of a number of important algorithms, but basically the idea is that a page handles a query when it either contains information relevant to the query, or points the reader to a resource that contains such information [294]. One approach is to look at the hyperlink structures that provide the context for webpages, and to try to deduce measures of relevance from those structures.

So, for instance, the simple Boolean model calculates the number of query terms that appear in the document, which can rank pages based on conjunctive queries, or transformations of disjunctions or negations into conjunctions. Then it is a reasonably logical step to use a recursive spreading activation algorithm to propagate the query, by looking for the query terms in neighbouring documents, reducing significance coefficients as the resources examined get further away from the original page [294].

On the most-cited model, a page is assigned a score which is the sum of the number of the query words contained in those pages which link to it. So the most-cited model finds authorities rather than hubs, although simple relevance (without attempting to privilege authorities over hubs) can also be generated using a spreading activation algorithm [76].

Beyond simple hyperlink connectivity, more sophisticated measures are based on the vector space model on which documents and queries are seen as vectors [76]. So, for example, TFxIDF gives a relevance score to a document based on the sum of weights of the query terms normalised by a Euclidian vector length of the document; weights of terms are calculated as the cross-product of Term Frequencies (TF) and Inverse Document Frequencies (IDF). A TF is a measure of the frequency of a term's occurrence in a document, while the IDF is a measure of the number of linked documents containing the term [180].

TFxIDF fails to take into account the important information provided by a page's hyperlink connections [47], but even including such information in a broader algorithm doesn't outperform TFxIDF by a huge distance [294, 76].

Another obvious measure of relevance in an e-commerce or e-publishing environment is to measure the number of downloads per visit. Such patterns of usage and acquisition can be studied to produce a map or trail of the way that knowledge is being transferred to and used by a user community. Experiments along these lines have shown that significant changes often happen very abruptly, alongside related events such as the creation of a link to the resource from an external site, or some discussion of the site by an external commentator [9].

The hyperlink structure in which a webpage finds its context is also informative about quality proxies. If there is a link from one page to another, that can be read as an endorsement of the second paper by the first. That is a defeasible hypothesis that depends to a large extent on the behaviour of the people who actually create webpages – it turns out that a large number of links are indeed endorsing other documents to some degree, even if only as an alternative source of information on the same topic. The mathematical measure is firmly embedded in the contingent sociology of the Web. Furthermore, such methods can be applied to multimedia items on the Web which may not contain any particularly interesting text on which to search, as for example with the pictorial retrieval system PicASHOW [182].

There are two main techniques for extracting quality information from hyperlink structures [76]. *Co-citation*-based methods are based on the insight that links to or from a page are likely to connote some kind of similarity. If two pages both point to a third page, then the first two pages may well share a topic of interest; if a page points to two other pages, then the latter two may also share a topic. *Random walk*-based methods use the model of the Web as a graph with pages as nodes and links as directed edges (see Section 4.1.2 above) and develop probability statistics based on random walks around it. Measures of quality of a page come out of such methods by measuring the quality of the other pages it is connected to, and filtering by the degree of those

connections. Together with relevance metrics, quality metrics can then rank the results of searches [76].

The most famous quality measure is PageRank [221], discussed earlier, which builds on the intuition that a page which is cited by many other pages is likely to be of significant quality. The insight of PageRank is that the obvious way to subvert that model is to set up a load of dummy pages to cite the page which one wanted to boost. But if a page is cited by many other pages *which themselves have a high PageRank*, then it is likely to be of high quality. The PageRank method has another intuitive characterisation that at first sight seems to have nothing to do with quality: it is the probability that a random surfer will reach the page [47]. The value of these measures is reflected in the success of Google in terms of longevity, market value and share of the search engine market. Further, other measures of quality exploit the idea of random walks [181], sometimes explicitly extending the ideas underlying PageRank [235].

A related idea is Kleinberg's HITS algorithm, based on the idea of *impact factors* from bibliometrics [171]. The original explanation of impact factors for academic journals was that one could look at the number of citations to a journal in the context of a discipline as a whole. One can then model the *influence weight* of a journal as a function of the influence weights of citing journals and the fraction of the citations of those citing journals that cite the journal in question. Analogous reasoning establishes an algorithm for measuring webpage quality, both in terms of its authority value and its hub value.

Patterns of usage can be characterised independently of measures of quality or relevance. What is the probability of a document being accessed within a particular time? What is the expected time before the next access of that document? Knowing the answers to such questions allows the identification of pages, resources and documents that are likely to be accessed frequently, in which case they can be *prefetched*, or made more available to users. Prefetching can be performed on the user's behalf, based on his or her particular use profile, or by a server based on statistics about use patterns in the population as a whole. Another application of such statistics is the development of adaptive

websites, where the presentation of material and the intra-site hyperlink structure can be varied automatically on the basis of the site's learning from previous usage [227]. Variables relating to usage patterns can be delved out of server logs containing the time and the URI of an access request, together with models of how the future probabilities depend on past usage [76].

5.4.4 Trust and reputation

Quality and significance are related to the reception that a page receives from a reader; the reader's beliefs about a page are inherently more subjective than the metrics outlined above. These subjective beliefs tend to be gathered under the heading *trust*. We have already seen the tendency for authorities and hubs to appear as the focus of cyber-communities. Such sites are in important ways *trusted*: authorities are trusted by other webpage authors to contain reliable information, while (successful) hubs are trusted by users to point to places where reliable information can be obtained.

Trust is, of course, an important factor in the development of the Web, in any number of fields. Scientific or academic papers are trusted to report correct results. Authors of pages are trusted to be who they say they are. Web services are trusted to do what they say they will do without damage to others. E-commerce sites are trusted to make proper use of credit card details, to send the goods ordered, and to keep data secure. The architecture of the Web, which explicitly facilitates anonymity and accurate copying, makes trust a particularly important issue.

Studying trust online is particularly difficult because of the multiple contexts in which online interactions take place. A recent survey [116] discovered that studies often failed to distinguish between trust, the causes of trust and the antecedents of trustworthiness. Trust is variously defined as 'confident expectation', 'a willingness to be vulnerable', 'a general positive attitude'. Trust in systems and trust in individuals are assimilated as if this is unproblematic. Focused empirical studies often rigorously and quite properly define their terms, but definitions are rarely common across studies, and so comparison is hard if not

impossible, and sometimes the rigorously-defined construct is barely recognisable as (folk-psychological) trust [215]. Trust is also not a static phenomenon, it is dynamic; there is often a period of time during which trust in a site is built up. Web users at different levels of experience also have different typical levels of trust [85].

All of this strongly implies that trust cannot be produced by creating the right tools and technologies – even if it was automatically a good thing to produce trust, which it is not (the key aim is to engineer a causal link between trust and trustworthiness). Trust will not magically appear online. Just as people will not automatically follow codes of conduct, others will not automatically assume that people follow codes of conduct. And because trust is not only a private good but a public one, people will always be able to 'free ride' on others' good behaviour [56].

There are two levels of significance with respect to the promulgation of trust across the Web which demand different approaches. First there is the level of the system as a whole, where one tries to verify that the rules governing the interaction force all the actors to be honest. The main strategy at this system level is to provide the infrastructure to ensure security, for instance with the use of certification schemes [230] or privacy-enhancing technologies [234], and takes the Hobbesian route to deterring immoral behaviour – it makes it too costly to perform, for one reason or another. For that reason, such mechanisms are strongly associated with issues to do with Web Governance.

Secondly, there is the level of the individual, at which one hopes that one's interactive partners or opponents are honest, reciprocative and rule-following. Here one tends to rely on feedback on behaviour; somehow a Web user gains a *reputation*. A reputation is a key element to trust, as it presents a sketch of the trustee abstracted (independently) from its history [205]. Based on history as it is, a reputation does not and cannot bind future behaviour; it does not therefore remove risk. Based on a wide range of judgements, most of which are subjective or have a strong subjective element, the aggregating function of a reputation is meant to smooth out particular rogue opinions or singular events.

Several lines of research are important to understanding how best to collect and understand reputations (cf. [239]). What methods will enable ratings to be gathered that define the trustworthiness of a Web user? How should such ratings be aggregated? How should we reason over these aggregations? And how should they be publicised? And as with many questions about the Web, what is the trade-off between a carefully-honed and accurate system that may be expensive to use, and a rough-and-ready system that has utility for 90% of purposes, which is trivial to use and has a large buy-in from a big user base.

eBay's reputation and feedback mechanism [86], where ratings are crude $+1$ or -1 values summed together with textual annotations, is of course the best example of a well-known reputation mechanism. Its reliability is open to challenge: some buyers don't return ratings; biases may occur (depending on whether good or bad experiences are more likely to be reported); bootstrapping reputation, before one has interacted at all, may be hard; one can imagine ways of manipulating the process (cf. [242]). On the other hand, the commercial success of eBay is self-evident, and the actual amount of fraud on eBay, despite some well-publicised cases, doesn't seem to be particularly large. Like Google, is this a case where a simple, scalable system seems to do very well?

A related and complex issue is that of finding metrics to measure trust for individual ratings and algorithms for sensible aggregations. Most metrics involve some score between $+1$ and -1, usually a real number. Two obvious issues emerge. Firstly, since our trust/distrust is rarely perfect, how should one choose a particular number? And secondly, how should be distinguish between two possible interpretations of 0, which could mean 'I have no experience with this person, so have no opinion', or 'I have experience, but I am neutral about him or her'. Furthermore, there are several sources of information about trust that seem to be important in making judgements: for instance, the previous interactions one has had with a trustee; reports of witnesses; certificates; and the role that the trustee is playing [152]. And, depending on requirements, one may prefer a trust value calculated on the basis of some objective or Archimedean perspective, or on the other hand a value calculated in the context of one's own preferences, beliefs and interests (and therefore a trustee's value could vary from enquirer to

enquirer). Evaluation of trust metrics is inevitably tricky, though not impossible if enough context can be provided to make the evaluation meaningful (cf. [114]).

Trust being a second-order good, hard to quantify, and task-relative, it would seem that all metrics would have to be approximate and would depend on what would work best; that is another argument for the relatively crude eBay approach. Having said that, there are systems, such as REGRET [246], which allow users to give their ratings richer content by annotating them; aggregation is performed by fuzzy reasoning.

Sometimes a quantitative metric would be inappropriate. For instance, when assessing information sources, it may be that a user really needs to see annotations and analysis by other users. The patterns of usage of information are certainly hard to quantify, and furthermore may be contradictory or incomplete; in that case it may be that, in compact, well-understood domains at least, semantic markup of documents may be the most helpful way forward [111]. The question then is how best to exploit the extra expressivity thereby gained: is it worth investing in formal languages, or a combined formal/semi-formal/informal approach? Nevertheless, real stories by real users, although they need high bandwidth, are often extremely informative.

The testimony of others, however gathered, represented or aggregated, is clearly of importance to the development and sustaining of reliable trust. The structure of the Web has proved suggestive in this field, in that the very Web-like structure that gets you to an arbitrary webpage in the World Wide Web can also get you quickly to the testimony of someone you don't know in a Web of Trust. As long as people store their experiences in a reliable manner, then they can be leveraged by other users by using aggregation algorithms [e.g. 243].

The requirement for such systems is that there is some information somewhere where people have described their beliefs about others, and have linked that information into the Web of Trust somehow. Once the information is available, it can be used to help determine a reputation. Such applications are beginning to emerge; one of the most prominent is FOAF [45] – http://www.foaf-project.org/), an RDF/OWL-based ontology which has been extended with a

vocabulary for describing one's relationships with and opinions of friends [115].

Trust, as has often been pointed out, is not transitive (that is, if A trusts B and B trusts C, it does not follow that A trusts C). That would seem to undercut Web of Trust approaches. However, if A trusts B, B trusts C and B recommends C to A, then that is a *reason* for A to trust C. The chain will break down eventually, but not necessarily immediately, and it may degrade gracefully. So as long as the notion of degradation is built into the generation of measures of trust based on Web of Trust approaches, then it would still be possible to model or generate trust based on eyewitness reports or stored opinions [115]. It has been argued that the expressivity of the Semantic Web is required to ensure that the aggregation of trust information is not merely heuristic in nature; it is the *content* of the attributions of trustworthiness or otherwise that counts. Once someone publishes a file which says who they know and how much they trust them, that social information can be processed without intermediaries [115, 243].

Future work in understanding how trust information can be extracted out of Web-like structures is a central topic in the exploration of social networks and their Web-like representations. Richardson et al have shown that path algebra and probabilistic interpretations of the exploration of the graph of a Web of Trust are nearly identical [243]; can this result be used as a method of ranking pages in Web searches? And all methods that exploit a Web of Trust simplify the attribution of trust; can methods be extended to include multi-valued beliefs and other data (such as metadata about provenance)? Given the importance of content to the mapping of Webs of trust, then it may well be that trust-generating techniques could play a similar role with the Semantic Web as algorithms such as PageRank, which extract information from uninterpreted link structures, play in the WWW.

5.4.5 Trust (II): Mechanising proof

There is, finally, a sociological coda related to trust: do we trust the machines and automated processes that are put under way when we work or play on the Web? It has been argued that culturally we now

deal with two notions of proof. In one view, as Wittgenstein argued, a proof is a picture which stands in need of ratification, which it gets when we work through it [292]; it convinces us. It explains and demonstrates the truth of the proved proposition simultaneously.

The other type of proof is mechanical and algorithmic; this may be more reliable than proof-as-picture, but to be accepted requires it to be taken on trust that the steps in the proof be done correctly. Trust is required (a) because the proof may be unsurveyable, and (b) even if not it is not efficient or cost-effective to check each mechanical proof by hand. Wittgenstein did not live to see complex mechanised proof become commonplace, but he did devote time to thinking about the implications, within his (at the time unusual) view of mathematics as an activity, and was careful to distinguish between proof-as-picture and mechanical proof. He concluded that our decisions to trust mechanical proofs are voluntarily and that their results are not forced upon us [292].

When extensive and complex mechanical proof appeared on the scene, the dilemmas that Wittgenstein predicted followed. For example the possibility of formal proof of the correctness of a program was debated in a couple of well-known and controversial articles. DeMillo et al claimed that proof-as-picture was required for systems to be (socially) usable, but that machines could not provide them [73]. Fetzer argued that there was a persistent confusion between two types of mechanical proof, one being a sequence of logical formulae where each formula is either an axiom of derived from the formulae above by truth-preserving rules, and the other being made by a machine [100]. Either way, the articles, and the fierce response to them, showed that the notion of automated proof was controversial.

Nowadays, many more aspects of daily lives (financial, health and safety, functioning of utilities) are under the aegis of automatic systems. And when the Web takes on more of the user's routine information processing tasks (as with the SW), the need for human trust in the mechanised system is all the greater. Much of that trust is an unpredictable function of experience [85], and we cannot obviate the need for trust in collective human judgement as well as in the machines themselves [192]. The relationship between trust in our collective selves

and trust in the hardware and software is a hard one to disentangle, and yet the development of the Web will depend crucially on it.

5.4.6 Web morality and conventional aspects of Web use

Moral and ethical questions are a necessary part of the Web Science agenda. They are necessary for our understanding of how the Web works, and, no less important, how the Web can grow.

The simplicity of the relationship between URIs and particular Web resources is key to leveraging the information space. Attempts to subvert this relationship can be very undermining of the Web and Semantic Web. Threats to that structure will undermine the link between the URI and what is displayed on the screen, and the more complex that the engineering gets, the harder it will be to detect such subversion.

The Web is a deliberately decentralised structure. The flip side of that is that there is no authority to enforce good behaviour. Although it is certainly the case that many types of behaviour essential for the Web to work (meaning, convention, commitment) are understandable from the point of view of rational self-interest [261], if we assume there are payoffs to bad behaviour, either of commission (opportunities to gain by cheating) or omission (failure to maintain a website satisfactorily), then self-interested rationality cannot *entirely* explain how such cooperative behaviour gets off the ground [144]. However far such analyses go, there is a deep non-rational element to such behaviour [254]; people must behave *well*.

There are plenty of texts about good behaviour, of course. The aim of this text is not to stake a claim to any of that territory. What counts in Web Science is the way that the engineering, the connection between URIs and what is displayed on the screen, depends on particular conventions of behaviour that is at some level altruistic. There may be things to say about sanctions to enforce such good behaviour (see Section 6), but it is not the place of a science of the Web to work out ways of providing moral leadership, or of working out the sometimes difficult conflicts that the desire to act morally often throws up. However there is a role for Web Science to determine what engineering

practices are important, and how they relate to people's willingness to behave in a cooperative fashion. Such analysis can lead to codes of behaviour that may not be enforceable but which in a sense define moral behaviour in the Web context. Morality and engineering turn out to be linked.

Let's follow the example of the connection between a URI and what it points to in detail. Unfortunately, as anyone who has had to maintain a Website will know, over time pressure undermining the connection builds up. Some pressure is caused by genuine engineering difficulties, some pressure is merely temptation or sloth. But the Web will function better if URIs don't change, if they always point to the same document (which of course may be updated periodically).

The number of working links actually declines quite rapidly. An experiment mentioned earlier crawled 150m webpages for 11 weeks, and by the 9th week the experimenters had lost access to over 10% of those pages (about 4% had disappeared within the first week). About 3% returned 4XX errors, most of those 404 errors (not found), and most of the rest 403s (forbidden). About 3% of the pages were blocked by Web servers' robots.txt files that detected and repelled Web crawlers. 2–3% of the failures were network-related, such as DNS lookup failures, refused connections or TCP timeouts, while about 2% were 3XX errors, indicating a page had moved. The .net and .com domains were apparently the worst offenders [99].

Avoiding changing URIs is easier said than done. For instance, when a website is reorganised, the temptation is to provide a neat new rational(ised) set of URIs expressing the new organisational philosophy. This is tempting, but ultimately unwise. Dangling links are frustrating, and actually do a lot to undermine trust in websites and companies (a functioning, well-presented and professional-looking website being an important reinforcer of online trust – cf. [116]). But given that all references to URIs by interested parties are 'out of date', in that they are records, stored in people's favourites lists, scribbled on paper or explicit links from other sites, of discoveries made in the past, they cannot be updated easily [27].

This is partly a question of style. [27] includes a set of suggestions about what *not* to include in naming directories and files: authors'

names, subjects, status, access rights, etc. All of the latter can seem quite sensible as filenames, but over the timescale of the Web these can change, which either creates pressure to alter the URI or renders the filename misleading (i.e. worse than meaningless). This means that producing URIs needs rather more thought than one would otherwise imagine, in that the webmaster needs to think about how to present a suite of information, and organise it, in such a way as to make sense in the future – at least the medium term. This is a real cost, but if the Web is to function well, most if not all webmasters must follow such conventions.

This is an example of the way morality hits engineering on the Web. Unlike the building of a complex artefact such as an aeroplane engine or ship, the individual 'workers' on the Web have not abrogated decision rights via contract. On the Web, everyone is a volunteer. But there are obligations, duties that one incurs by being online because of the cooperative nature of the Web, and meeting these obligations is part of the task of creating the important invariants in the Web experience. Another example, on the personal level, is keeping content up to date and accurate.

Socially, it is important to identify and try, where possible, to engineer out harmful behaviour (harmful both to individuals and to the Web as a whole) such as phishing, or hoaxing PageRank and other search engine algorithms. There will be no truly engineering solution to such behaviour; it occurs within a given Web context, and those who indulge in it will always be tempted to work around any current block. But codes of conduct and other types of discussion about the Web can create consensus about what constitutes online duty and what constitutes bad behaviour (context is important: why is spam a serious irritant, and junk mail relatively minor?) and, consequently, about what behaviours should be legitimated, what mandated, and what related functionality architectures might be expected to provide. The close link online between engineering and morality is unusual if not unique. The fleshing out of these obligations is a remarkable aspect of our understanding of the Web, and in our final substantive section we look at some of the issues that it raises in more detail.

6

Web Governance, Security and Standards

Knowledge, goes the cliché, is power. The Web, by dramatically shifting the structures underlying knowledge and its accessibility, has altered power structures in ways that are ultimately unpredictable. The timeless truths of politics and society haven't been changed by the advent of the Web [217], but their context has. Power has shifted, and this raises the question of Web governance. How should things be regulated to ensure the steady and fruitful development of the Web?

We have already seen, in Section 5.4.6, that regulation cannot be the answer to everything. The *general* problem of Web governance is that with a decentralised structure it is hard to enforce standards, and with a very large number of untrained or relatively uninterested users things have to be kept very simple. But that simplicity can't be allowed to stand in the way of people being able to formulate policies about access and control, and to implement them. It is arguable that the relative lack of sophisticated information controls have hindered the growth of the Web by making people reluctant to make information available, and thus to share it with the community [287]; security and privacy are extremely important issues too.

Different information providers, with different policies governing control of information (or indeed no policies at all), will have problems sharing, and the problem will get worse if sharing is done on the coarse level of webpages, documents or websites, rather than at the finer-grained level of the individual piece of information. On the other hand, it is equally true that there are a number of platforms, protocols and architectures that facilitate information security, but which are not widely used. And an added constraint is that infrastructure has to enable security, privacy and trust without bothering users with constant information or requests for permissions. The governance of the Web cannot be neglected by Web Science. We begin our discussion of aspects of this space with the processes of standards-setting and policy-making.

6.1 Standards and policies

Standard-setting allows industry-wide cost savings thanks to economies of scale (cf. e.g. [281]), and so is generally speaking a good thing. But there are potential pitfalls [36]. It may be that one or two large firms have the capability in an industry to dominate standards, and ensure that smaller competitors and suppliers follow. Market leaders can use such standards to stay one or two steps ahead of the pack. Standards wars can be wasteful of R&D effort (cf. the recent battles over the next generation of DVD formats). Negotiated standards, where everyone prefers a standard to no standard, are likely to produce the best outcomes in an industry, and the existence of effective bodies, perceived to be neutral, whose only agenda is an engineering one, is an important aspect of Web governance.

In the case of the Web, standards are needed to ensure the preservation of its essential architectural properties, combined with decentralisation, flexibility and usability, in a sphere where the social aspects of use are not yet fixed. Information-sharing has traditionally been limited, and embedded within well-understood contexts. So, for instance, sharing a photograph has traditionally involved handing over a physical object. The trajectory of such an object is relatively easily traceable. Misuse of the object is relatively detectable. And even if the actual

misuser cannot be found, culpable individuals (i.e. the person who lent the photograph without permission) can be. Digital technologies have changed all that. Sharing a digital photograph facilitates massive copying and dissemination with little recourse to the user, even if it is discovered.

Standards and policies designed to make good behaviour easier and more likely are therefore required. Such policies, typically, will specify who can use or modify resources, and under what conditions. Policy awareness involves ensuring users have accessible and understandable views of policies associated with particular Web resources, which will not only support good behaviour but make it possible to identify breaches and thereby root out bad behaviour. The space for policy aware infrastructure will be in the deployment of the upper layers of the Semantic Web, as shown in Figure 3.2. Rules should be deployable which will enable the scalable production and exchange of proofs of rights of access [287].

Policy awareness, because of the particular context of the Web, will have to be markedly different from current approaches to information security and access control, which exploit mechanisms that require coordination and costly maintenance (e.g. PKI systems), and which therefore are over-prescriptive for general use on the Web. Even routine password-controlled access can be irksome. Weitzner et al describe the dilemma of someone wanting temporary access to restricted material. Raising that person's security grade risks allows him or her to see other restricted material, while declassifying the material risks allows others access to it [287].

The Web requires creative description of security measures, rather than prescriptions and mechanisms, and a number of approaches have been developed for framing policies. Ponder is an expressive policy description language for distributed systems, but being mainly syntactically-based may not work well in a more semantically-enabled future [68]. KAoS, a policy representation language based on OWL [275], and Rei, which allows agents to control access using policies described using OWL ontologies [161], also make interesting suggestions about access control and information sharing in distributed systems of agents or Web services. Work on the Policy Aware Web goes

beyond this agent/service-based paradigm; a beginning has been made on infrastructures appropriate to the decentralised and democratic Web, but much remains to be done (for example, on appropriate user interfaces) to ensure that transparency and accountability of information use are properly in place.

6.2 Copyright issues

As the Web is an information space, a vital area is that of copyright and intellectual property. Copyrights protect the expression of an idea, and so are narrow – they don't prevent others releasing, say, novels with similar storylines to a novel currently under copyright – and are aimed to protect an author's, musician's or other creative person's distinctive contribution. The narrowness makes it hard to use copyright law in the commercial software arena, so for instance the US Supreme Court upheld Borland's appeal against Lotus after the latter sued the former for 'borrowing' features of Lotus 1-2-3's interface. There are now extensive rights in both the US and Europe allowing reverse engineering and copying to produce compatibility, in the public interest [247].

Databases, treated as compilations, have been in receipt of the same protection as literary works (i.e. protected for 50 years after the creation or 70 years after the death of the creator in the UK), but following an EU directive in the late 1990s, a database is protected for 15 years following its last major change. The selection of information and its arrangement must amount to an intellectual effort to obtain, verify or present it. There have as yet been very few cases brought to establish precedents, but given the quantity of the deep Web that is contained in databases, and the aims of the Semantic Web community to bring together distributed information from a range of relational databases, it is quite likely that database rights will become the subject of increasingly searching debate in the future [132]. More generally a new European directive (2003/98/EC, http://www.ec-gis.org/document.cfm?id=486&db=document) on Public Sector Information has come into force. One of its objectives is to expedite the publication of and access to the considerable amounts of data collected by governments in their various functions. In the UK this has

led to the creation recently of an Office of Public Sector Information (www.opsi.gov.uk) – they are taking a close look at whether the SW is an appropriate vehicle for fulfilling their obligations.

Copyright is currently the focus for a major argument in the field of intellectual property law. Some stakeholders point out that digital technology, and the connectivity of the Web, have between them made piracy very straightforward – copying and distribution are the simplest things in the world, and so they support the development of technologies and legal instruments to prevent or limit unauthorised reproduction. Others point out that the power of the Web comes precisely from the serendipitous reuse of content, and that most uses of information, particularly in the context of the Web, are harmless and desirable, most of all in academe [131]. The argument turns on whether creativity is more likely to be stifled by the loss of incentives for authors whose copyright becomes worthless, or the shrinking of the commons and the public domain [93]. Lawrence Lessig has argued for the idea of a 'creative commons' (http://creativecommons.org/), which is intended to offer a flexible range of protections for works that do not stifle openness. Metadata is attached to works effectively waiving some or all of the rights that copyright law provides the author [187].

There are similar divisive arguments about patents, which give inventors a twenty year monopoly of the use of a new, useful and non-obvious discovery, but these arguments require as much discussion of institutions and governmental procedures, and the wider economics of intellectual property. Patents (and trade secrets) are reviewed in [93].

6.3 Transgressive behaviour

In many cases, understanding how transgression can take place will suggest methods for undermining the transgression, but one must always be prepared for an arms race. So, for instance, so-called 'spamdexing', or the placing of particular keywords in a document so as to increase the probability of a search engine alighting upon it whether or not the contents are irrelevant, becomes less attractive as a policy for ensuring visibility of a webpage if quality measures focus on hyperlink structures rather than the content of the page [76].

As the most prominent example of the subject of an arms race, Google's PageRank algorithm [221] is a quality/relevance measure of great renown. So influential is Google on patterns of Web use, PageRank has to operate in a world where many agents are actively trying to subvert it.

Studies of the PageRank algorithm, or algorithms of that style, have often picked out the possibilities of free-riding on its work to promote spam [37], [178], and there are many 'how to' papers for would-be spammers. Former research director at Google Monika Henzinger identifies this as an important challenge for Google [137]. As long as there is advantage to be gained from appearing high up in lists of retrieved pages, the arms race will go on, and it is hard to imagine how techniques for spamming search engines could be made illegal – after all many of them simply exploit the linking or keyword mechanisms that make the Web so powerful.

6.4 Privacy and identity

Another issue, like spam, that worries people very much is that of privacy. The Web allows unprecedented data collection in quantities that is creating a backlash of users who are either deeply worried about the loss of privacy, or alternatively find the effort of working round such issues tedious [85], [234]. Information is often used for purposes different from those which may have been given as the reason for collection. And data security is all too often treated as a side-issue by firms, a fact highlighted in 2005 when it was discovered that leaks from various businesses had exposed the personal information of more than 50,000,000 people. America's resistance to privacy legislation has meant that such exposure often goes undetected, though a pioneering law in the State of California obligated firms to inform those whose data had leaked, and the scandal was uncovered. At the time of writing, privacy laws were gaining support at all levels of American society and government, and Microsoft had reversed its position and supported a federal privacy law [265]. Nevertheless, a recent survey reported that 59% of computing experts who responded to a questionnaire expected online surveillance to increase over the next few years [103].

The W3C promotes the Platform for Privacy Preferences (P3P) to enhance user control by allowing better presentation of privacy policies, and therefore allowing users to customise them more easily [67]. P3P is a standard that allows a common view to be taken of privacy by various different actors. Regulation of privacy is clearly on the political agenda, and arguably needs to be. However, it is also the case that sensitivity to personal requirements and preferences requires clever interfaces and useful tools and techniques for inexperienced or relatively uninterested users to protect themselves [286]. Further, P3P and similar approaches carry no enforcement mechanism when violated.

It may well be that the notion of privacy as it has been traditionally understood in post-Enlightenment polities will be too hard to protect in an era when the Web is used for so many transactions, contains so much information, and enables so much useful information to be gleaned from Web users without their knowledge. Some reasons why the digital cyberworld seems to be so inimical to privacy include: the potential longevity of stored information; the ease of copying and transfer; the accuracy of copying and transfer; effective search mechanisms; the power of amalgamated databases; the difficulties of suppressing information; the fluidity of identity and anonymity that the Web provides; lack of centralisation; the dearth of arenas for well-publicised error correction; the difficulty in identifying breaches of privacy; the difficulty of tracing culprits; the comprehensiveness of the Web's coverage of our lives; its pervasiveness in our lives; digital information's independence of medium; the compact architectures on which information is stored; the strange affinity between the Web and subterranean behaviour. No doubt there are many more reasons; compare them to other information storage media, such as paper, and it can be seen how much greater a threat to privacy the Web is. It used to be the case that even when stored, information was in practice nearly impossible to find (for example, in a large paper-based filing system that has evolved piecemeal over the years). In our digital age, this phenomenon, which David Brin has called 'practical obscurity' [46], is a thing of the past.

But that needn't be the end of the matter. Perhaps the focus should be on the definition of what constitutes misuse of information, and

perhaps we should move towards standards that promote accountability of information users, and transparency in the way information is used. We need to understand how regulation and platforms that enable user control interleave with ordinary working or private lives. After all, there are currently many platforms in place to aid information security and protection of intellectual property, but people tend not to use them. As we noted above, we haven't yet struck the ideal balance between bothering people with queries on the one hand, and allowing information to become dangerously insecure on the other.

A related issue to privacy is that of identity and authentication. As more automated systems depend on one being able to prove identity (in order, for example, to get access to resources), the need for authentication increases. The fluidity of identity has often been cited as one of the most important attractors of the Internet ("no-one knows you're a dog" – cf. [186, 215]), but identity assurance systems needn't necessarily compromise that. In particular, in the absence of biometric standards we might assume that 'identification' and 'authentication' are more or less synonymous – a person is identified via something he or she owns (e.g. a smart card, door key, household bill), something known (e.g. password, answer to a specific question, PIN number), or some personal characteristic. Many systems do not include a personal characteristic in the loop, and therefore equate the individual with the authentication method; they assume that an initial, accurate authentication took place, and then rely inductively on the assumption [230]. Whatever the drawbacks of such an assumption are – and from the point of view of security they are many – they do at least generate a relative standard of identity rather than an absolute one, and are therefore less intrusive.

The obvious point to make about identification mechanisms is that the easier they are to use, and therefore the more suitable for the heterogeneous user base of the Web, the simpler they are to compromise. Fixed passwords are familiar and easy to administer, but vulnerable to simple attacks; on the other hand public key-based identification protocols are cryptographically powerful (and indeed computationally cheap and still relatively simple), but they generally need something like a hardware token as well as supporting infrastructure [230].

6.5 The economics of information and communication

The Web is not only a political space; it is also an economic space, because knowledge has a value [209], although as with politics the new online environment doesn't entail that economics textbooks be torn up. For instance, the general structure of an information industry – with relatively large fixed costs (to discover or acquire the information) and negligible marginal costs (each copy of the information is trivial to create) – suggests that they are natural monopolies; once the fixed costs have been undertaken by a firm, then they can always price new firms out of the market as long as they can hamper the other firms' acquisition of the requisite information. Work needs to be done to determine how far this sketch of the economic situation is true; for example, it seems that online firms have competed for market share, which has led to remarkably low online prices. To the extent that the sketch is true, however, the question of regulation of those natural monopolies must raise its head (cf. [281]).

Search may be an issue. Where there are bottlenecks, there are monopoly opportunities. Search can be regarded as an important bottleneck on the Web (cf. [25]). The major search companies face increasing scrutiny (in common with other firms in the field of computing) as they have to deal with the problems of internationalisation and conflicting political requirements, perhaps most notoriously in China [16].

6.6 A liberal hegemony?

A final point briefly worth making is that the Web is a space designed to let information flow, and to create opportunities for cooperation and collaboration. It is worth asking why freer information flow is a good thing, and the answers are pretty straightforward. It is good to have the freedom to express oneself in order that one can pursue one's own autonomous and authentic projects. Unhindered criticism of governments and other power centres tends to lead to better governance; information drives democracy. Both of these reasons have their roots in a liberal, individualistic view of the world, in the tradition of Locke, Mill and Rawls. Perhaps the Web is a liberal artefact?

There is certainly opposition to the Web from many sources (most of these sources, it is fair to say, are more than happy to employ the Web as a tool for organisation, communication and dissemination). Many illiberal governments restrict their citizens' use of the Web, often using adaptations of firewall technology to create what is in effect a giant intranet within their borders. Even non-liberal democracies have something of a problem with the Web. For instance, the government of Singapore has a relatively light touch in its regulation of the Internet, but still blocks 100 or so pornographic sites, requires political and religious websites to be registered and licensed with the Singapore Broadcasting Authority, and bans election activity on the Web during election campaigns [197], even though it has a powerful vision of a knowledge-based economy and is one of the most enthusiastic governments in the world with respect to IT [273].

In the realm of non-governmental activity, the Web has also been seen as an agent of globalisation, and so views of the Web have been conditioned by authors' political views about that trend. Many see the Internet as a wonderful anarchistic paradise while the Web, with its slick websites and mass appeal, has destroyed all that and normalised the online world. Online is just as grim and unjust, for such writers, as offline [241]. Marketing has replaced democracy. In these discourses, neologisms such as 'cyberhegemony' and 'cyberdependency' abound [226].

For the Web to be a contributor to global well-being its developers have to pick their way through a number of tricky debates such as this; it is essential that the Web does not become a global monoculture, while also avoiding the alternative of decomposing into several cultish mini-webs with little or no connectivity in between. The balance of respect for others' points of view and proper defence of one's own has always been a difficult one to strike in any sphere of human activity. At the moment, the Web surprises us with the fruitfulness of its connectivity. It is important that this is retained [30]. It may be that the fractal structure of the Web, if it can be nurtured, will be part of a solution [29]. We also need to understand the way that the Web is used in developing nations, rather than focusing on the Western democracies,

in order to ensure that it can serve as wide a set of constituencies as possible [83].

Given all these worries, it is perhaps unsurprising that the US government has recently come under pressure about the prominence of its role in Web governance, despite the obvious success of the Internet and the Web so far. The United Nations Working Group on Internet Governance's 2005 report made a number of recommendations that all stakeholders should be involved in Internet governance [288]. This might change the liberalism of the Web. The likely effects of this on the Web itself are unknown (cf. [274]).

7

Discussion and Conclusions

This text has enumerated a series of approaches to both understand and engineer the Web. We argue that these approaches can be organised into a framework and that such a framework constitutes a science for our discipline. In this science we need to investigate architecture and we need to understand and formulate our architectures at appropriate levels of abstraction. A Web Science will contain its own debates about appropriate methodology. It is unavoidably a combination of synthesis, analysis and governance – since the Web exists within a complex set of social and legal conventions.

We have argued at length that a move from a document-centric Web to a more thoroughgoing data Web is likely to require more by way of semantic technologies. Not least because of the fact that the transparent and unambiguous integration of heterogeneous data demands clear semantic description. The extent to which ontologies will provide a widespread mechanism to achieve this mediation was discussed. Whether ontologies or folksonomies, if we are to coordinate our Web of data then stable vocabularies of varying scale are an important element. In a web of data familiar problems of referential identity arise. When are two concepts the same? How are we to construct robust and

flexible naming schemes? How are we to account for the natural drift and evolution in our interpretation of the meaning of concepts?

Current trends in Web research will change the nature of the Web itself. Whether this is the emergence of Web services, new models of content sharing such as P2P, the demand for personalisation, widespread automatic Natural Language Processing or the emergence of mobile computing, each of these topics will be legitimate components of our Web Science.

We have also reviewed the various approaches that seek to analyse the Web as it is and as it may become. Here the need is for researchers in mathematics and physics, biology and economics to make common cause with engineers and computer scientists to help enrich our understanding of this huge decentralised information system. We have not said much about how understanding and analysing the Web could lead to important insights for other disciplines. But this is almost certainly going to be the case. Serious scientific collaboration is never a one way street.

We have spent time articulating the challenges that Web Science raises from a moral and societal viewpoint. We believe this to be indispensable. The Web perhaps more than any other recent human construct carries with it any number of issues including privacy and protection, access and diversity, control and freedom. Structures that we design, engineer and research, and findings that emerge through analysis, will often have strong societal implications. We are keen that the Web Science community is socially aware, informed and where necessary proactive.

Finally, we believe that the arguments about whether a science should be essentially analytic are sterile [34]. We require science to analyse and synthesise. We also suspect there is more art to science and science to art than is often acknowledged. We are more than happy to acknowledge that Web Science is an electic discipline. We also believe that it possesses some of the most challenging and intriguing questions of the 21st century.

Acknowledgements

Thanks are due to the participants in the workshop on Web Science held at the British Computer Society in London, 12th-13th September, 2005, for two days of stimulating discussion that helped shape our ideas of what the science of the Web consists in. As well as the present authors, the participants included Hal Abelson, Mark Ackerman, David de Roure, William Dutton, Joan Feigenbaum, Dieter Fensel, Carole Goble, Craig Knoblock, Ora Lassila, Robin Milner, Guus Schreiber, Henry Thompson, Yorick Wilks and Jonathan Zittrain. These participants of course are not responsible for the ideas put forward in this text, but we have tried to incorporate as many as possible of their ideas of the major issues for Web Science. Many thanks also to James Finlay, Susan Davies and Timothy Miles-Board.

During the writing of this text some of the authors were supported by the UK Engineering and Physical Sciences Research Council's Advanced Knowledge Technologies interdisciplinary research collaboration (grant number GR/N15764/01), and some of the work reported was conducted under the UK Economic and Social Research Council project 'Justice On-Line: Distributing Cyberspace Fairly' (award number RES-000-22-0563). We also thank the US National Science Foundation for their support of work in the Policy Aware Web and Transparent Accountable Datamining Initiative.

References

[1] M. Abadi and C. Fournet, "Mobile values, new names and secure communication," in *Proceedings of 28th ACM Symposium on Principles of Programming Languages (POPL '01)*, 2001. http://research.microsoft.com/users/fournet/papers/mobile-values-new-names-and-secure-communication.pdf.

[2] K. Aberer, P. Cudré-Mauroux, and M. Hauswirth, "Start making sense: The Chatty Web approach for global semantic agreement," *Journal of Web Semantics*, vol. 1, no. 1, http://www. websemanticsjournal.org/volume1/issue1/Abereretal2003/index.html, 2003.

[3] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," *2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, WWW05*, http://www.hpl.hp.com/research/idl/papers/politicalblogs/AdamicGlance BlogWWW.pdf, 2005.

[4] L. A. Adamic and A. B. Huberman, "The Web's hidden order," *Communications of the ACM*, vol. 44, no. 9, http://www.hpl.hp.com/research/papers/weborder.pdf, 2001.

[5] E. Adar and L. A. Adamic, "Tracking information epidemics in blogspace," *Web Intelligence 2005*, http://www.hpl.hp.com/research/idl/papers/blogs2/trackingblogepidemics.pdf, 2005.

[6] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[7] K. Ahmad, M. Tariq, B. Vrusias, and C. Handy, "Corpus-based thesaurus construction for image retrieval in specialist domains," in *Advances in Information Retrieval: Proceedings of the 25th European Conference on IR Research (ECIR 2004)*, (F. Sebastiani, ed.), pp. 502–510, Berlin: Springer, 2003.

[8] K. Ahmad, B. Vrusias, and M. Zhu, "Visualising an image collection?," *Proceedings of 9th International Conference on Information Visualisation (IV '05)*, pp. 268–274, 2005.

[9] J. Aizen, D. Huttenlocher, J. Kleinberg, and A. Novak, "Traffic-based feedback on the Web," *PNAS*, vol. 101, (April 6th 2004), http://www.pnas.org/cgi/reprint/101/suppl_1/5254, 2004.

[10] H. Alani, "Ontology construction from online ontologies," *Proceedings of WWW 2006*, http://www2006.org/programme/files/pdf/4013.pdf, 2006.

[11] H. Alani, S. Dasmahapatra, N. Gibbins, H. Glaser, S. Harris, Y. Kalfoglou, K. O'Hara, and N. Shadbolt, "Managing reference: Ensuring referential integrity of ontologies for the Semantic Web," in *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, (A. Gómez-Pérez and V. R. Benjamins, eds.), pp. 317–334, Berlin: Springer, 2002.

[12] H. Alani, S. Dasmahapatra, K. O'Hara, and N. Shadbolt, "Identifying communities of practice through ontology network analysis," *IEEE Intelligent Systems*, pp. 18–25, http://eprints.ecs.soton.ac.uk/7397/, March/April 2003.

[13] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the World-Wide Web," *Nature*, vol. 401, pp. 130–131, 1999.

[14] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, L. Ding, P. Kolari, A. P. Sheth, I. B. Arpinar, A. Joshi, and T. Finin, "Semantic analysis on social networks: Experiences in addressing the problem of conflict of interest detection," *Proceedings of WWW 2006*, http://www2006.org/programme/files/pdf/4068.pdf, 2006.

[15] A. R. Anderson and N. D. Belnap, *Entailment: The Logic of Relevance and Necessity vol.1*, Princeton: Princeton University Press, 1975.

[16] Anonymous, "Fuzzy maths," *The Economist*, 11th May 2006.

[17] G. Antoniou and F. van Harmelen, *A Semantic Web Primer*, Cambridge MA: MIT Press, 2004.

[18] A. Arasu, J. Novak, A. Tomkins, and J. Tomlin, "PageRank computation and the structure of the Web: Experiments and algorithms," *Proceedings of the 11th World Wide Web Conference*, 2002.

[19] K. Baclawski and T. Niu, *Ontologies for Bioinformatics*, Cambridge MA: MIT Press, 2005.

[20] S. Baluja, "Browsing on small screens: Recasting Web-page segmentation into an efficient machine learning framework," *Proceedings of WWW 2006*, http://www2006.org/programme/files/pdf/2502.pdf, 2006.

[21] A.-L. Barabási, "The physics of the Web," *Physics World*, http://physicsweb.org/articles/world/14/7/09, July 2001.

[22] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: The topology of the World Wide Web," *Physica A*, vol. 281, pp. 69–77, 2000.

[23] J. Basney, W. Nejdl, D. Olmedilla, V. Welch, and M. Winslett, "Negotiating trust on the Grid," *Proceedings of Dagstuhl Seminar on Semantic Grid: The Convergence of Technologies*, http://drops.dagstuhl.de/opus/volltexte/2005/387/pdf/05271.OlmedillaDaniel.Paper.387.pdf, 2005.

[24] T. Basten, *In Terms of Nets: System Design With Petri Nets and Process Algebra*, Ph.D. thesis, Eindhoven University of Technology, 1998.

[25] J. Battelle, *The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed Our Culture*, Boston: Nicholas Brealey Publishing, 2005.

[26] T. Berners-Lee, "The Myth of Names and Addresses," http://www.w3.org/DesignIssues/NameMyth.html, 1996.

[27] T. Berners-Lee, "Cool URIs Don't Change," http://www.w3.org/Provider/Style/URI, 1998.

[28] T. Berners-Lee, "Relational Databases on the Semantic Web," http://www.w3.org/Design Issues/RDB-RDF.html, 1998.

[29] T. Berners-Lee, "The Fractal Nature of the Web," http://www.w3.org/Design Issues/Fractal.html, 1998/2005.

[30] T. Berners-Lee, *Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor*, London: Texere Publishing, 1999.

[31] T. Berners-Lee, "What Do HTTP URIs Identify?," http://www.w3.org/DesignIssues/HTTP-URI.html, 2002/3.

[32] T. Berners-Lee, "Foreword," in *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential*, (D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, eds.), Cambridge MA: MIT Press, 2003.

[33] T. Berners-Lee, R. T. Fielding, and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax," http://www.gbiv.com/protocols/uri/rfc/rfc3986.html, 2005.

[34] T. Berners-Lee, W. Hall, J. Hendler, N. Shadbolt, and D. Weitzner, "Web Science," *Science*, vol. 313, August 11th 2006.

[35] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21, May 2001.

[36] S. Besen and J. Farrell, "Choosing how to compete: Strategies and tactics in standardization," *Journal of Economic Perspectives*, vol. 8, pp. 117–131, 1994.

[37] M. Bianchini, M. Gori, and F. Scarselli, "Inside PageRank," *ACM Transactions on Internet Technology*, vol. 5, no. 1, pp. 92–128, 2005.

[38] P. E. Black, ed., ch. Dictionary of Algorithms and Data Structures, *Levenshtein distance*. 2005. http://www.nist.gov/dads/HTML/Levenshtein.html.

[39] D. C. Blair, "Wittgenstein, language and information: 'Back to the rough ground'," in *Context: Nature, Impact and Role – 5th International Conference on Conceptions of Library and Information Science (CoLIS 2005)*, (F. Crestani and I. Ruthven, eds.), pp. 1–4, Berlin: Springer, 2005.

[40] R. A. Botafogo, E. Rivlin, and B. Shneiderman, "Structural analysis of hypertexts: Identifying hierarchies and useful metrics," *ACM Transactions on Information Systems*, vol. 10, no. 2, pp. 142–180, 1992.

[41] P. Bouquet, F. Giunchiglia, F. van Harmelen, L. Serafini, and H. Stuckenschmidt, "Contextualizing ontologies," *Journal of Web Semantics*, vol. 1, no. 4, pp. 325–343, 2004.

[42] R. J. Brachman and J. G. Schmolze, "An overview of the KL-ONE knowledge representation system," *Cognitive Science*, vol. 9, pp. 171–216, http://www.cogsci.rpi.edu/CSJarchive/1985v09/i02/p0171p0216/MAIN.PDF, 1985.

[43] B. Brewington and G. Cybenko, "How dynamic is the Web?," *Proceedings of the 9th World Wide Web Conference*, http://www9.org/w9cdrom/264/264.html, 2000.

[44] D. Brickley and R. V. Guha, eds., *RDF Vocabulary Description Language 1.0: RDF Schema*. 2004. http://www.w3.org/TR/rdf-schema/.

[45] D. Brickley and L. Miller, "FOAF Vocabulary Specification," http://xmlns.com/foaf/0.1/, 2005.

[46] D. Brin, *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?*, New York: Basic Books, 1998.

[47] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Proceedings of the 7th World Wide Web Conference*, 1998.

[48] A. Broder, S. Glassman, M. Manasse, and G. Zweig, "Syntactic clustering of the Web," *Proceedings of the 6th World Wide Web Conference*, 1997.

[49] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, S. Stata, A. Tomkins, and J. Wiener, "Graph structure in the Web," *Computer Networks*, vol. 33, no. 107, 2000.

[50] F. P. Brooks, *The Mythical Man-Month: Essays in Software Engineering 2nd Edition*, Boston: Addison Wesley Longman, 1995.

[51] R. A. Brooks, "How to build complete creatures rather than isolated cognitive simulators," in *Architectures for Intelligence*, (K. VanLehn, ed.), pp. 225–239, Hillsdale N.J.: Lawrence Erlbaum, 1991.

[52] A. Caldas, P. A. David, and O. Ormanidhi, "Digital Information Network Technologies, Organizational Performance and Productivity," Stanford Institute for Economic Policy Research discussion paper 05-11, http://siepr.stanford.edu/papers/pdf/05-11_summary.pdf., 2005.

[53] F. Carmagnola, F. Cena, C. Gena, and I. Torre, "A multidimensional approach for the semantic representation of taxonomies and rules in adaptive hypermedia systems," *Proceedings of the Workshop on Personalisation on the Semantic Web: PerSWeb '05*, pp. 5–14, http://www.win.tue.nl/persweb/full-proceedings.pdf, 2005.

[54] L. Carr, S. Bechhofer, C. Goble, and W. Hall, "Conceptual linking: Ontology-based open hypermedia," *Proceedings of 10th World Wide Web Conference*, http://www10.org/cdrom/papers/frame.html, 2001.

[55] E. Castronova, *Synthetic Worlds: The Business and Culture of Online Games*, Chicago: University of Chicago Press, 2005.

[56] J. Cave, "The economics of cyber trust between cyber partners," in *Trust and Crime in Information Societies*, (R. Mansell and B. S. Collins, eds.), pp. 380–427, Cheltenham: Edward Elgar, 2005.

[57] CCSDS, "Reference Model for an Open Archiving Information System (OAIS)," Consultative Committee for Space Data Systems, http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf, 2002.

[58] Cedars, "Cedars Guide to Preservation Metadata," http://www.leeds.ac.uk/cedars/guideto/metadata/guidetometadata.pdf, 2002.

[59] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger, "The origin of power laws in Internet topologies revisited," *IEEE Infocom*, 2002.

[60] P. C.-H. Cheng, "Diagrammatic knowledge acquisition: Elicitation, analysis and issues," in *Advances in Knowledge Acquisition: Proceedings of the 9th European Knowledge Acquisition Workshop*, (N. Shadbolt, K. O'Hara, and G. Schreiber, eds.), pp. 179–194, Berlin: Springer, 1996.

[61] T. Coates, M. Biddulph, P. Hammond, and M. Webb, "Reinventing radio: Enriching broadcast with social software," *O'Reilly Emerging Technology Conference*, http://conferences.oreillynet.com/cs/et2005/view/e_sess/5981, 2005.

[62] E. F. Codd, "A relational model of data for large shared data banks," *Communications of the ACM*, vol. 13, no. 6, pp. 377–387, 1970.

[63] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, "Resilience of the Internet to random breakdowns," *Phys Rev Lett*, vol. 85, http://www.wisdom.weizmann.ac.il/∼recohen/publications/internet_prl.pdf, 2000.

[64] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, "Breakdown of the Internet under intentional attack," *Phys Rev Lett*, vol. 86, http://www.wisdom.weizmann.ac.il/∼recohen/publications/attack_prl.pdf, 2001.

[65] D. Connolly, F. van Harmelen, Ian Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein, "DAML+OIL Reference Description," http://www.w3.org/TR/daml+oil-reference, 2001.

[66] L. Coyle, D. Doyle, and P. Cunningham, "Representing similarity for CBR in XML," in *Advances in Case-Based Reasoning: Proceedings of the 7th European Conference on Case-Based Reasoning*, (P. Funk and P. A. G. Calero, eds.), Berlin: Springer, 2004. https://www.cs.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-25.pdf.

[67] L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle, "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification," http://www.w3.org/TR/P3P/, 2002.

[68] N. C. Damianou, *A Policy Framework for Management of Distributed Systems*, Ph.D. thesis, Imperial College, London, 2002. http://www-dse.doc.ic.ac.uk/Research/policies/ponder/thesis-ncd.pdf.

[69] H. T. Dang, "Overview of DUC 2005," *Proceedings of DUC 2005*, http://www-nlpir.nist.gov/projects/duc/pubs/2005papers/OVERVIEW05.pdf, 2005.

[70] M. d'Aquin, J. Lieber, and A. Napoli, "Decentralized case-based reasoning for the Semantic Web," *Proceedings of the International Semantic Web Conference (ISWC 2005)*, http://www.loria.fr/equipes/orpailleur/Documents/daquin05b.pdf, 2005.

[71] W. Davies, "You Don't Know Me, But ...: Social Capital and Social Software," The Work Foundation, http://www.theworkfoundation.com/pdf/1843730103.pdf, 2003.

[72] D. C. De Roure, N. R. Jennings, and N. R. Shadbolt, "The Semantic Grid: Past, present and future," *Proceedings of the IEEE*, vol. 93, no. 3, pp. 669–681, 2005.

[73] R. A. DeMillo, R. J. Lipton, and A. J. Perlis, "Social processes and proofs of theorems and programs," *Proceedings of the 4th ACM Symposium on Principles of Programming Languages*, pp. 206–214, 1977.

[74] R. Denaux, L. Aroyo, and V. Dimitrova, "OWL-OLM: Interactive ontology-based elicitation of user models," *Proceedings of the Workshop on Personalisation on the Semantic Web: PerSWeb '05*, pp. 34–46, http://www.win.tue.nl/persweb/full-proceedings.pdf, 2005.

[75] A. Deutsch, L. Sui, and V. Vianu, "Specification and verification of data-driven Web services," *Proceedings of PODS '04*, 2004.

[76] D. Dhyani, W. K. Ng, and S. S. Bhowmick, "A survey of Web metrics," *ACM Computing Surveys*, vol. 34, no. 4, pp. 469–503, 2002.

[77] E. Dijkstra, "Go To statement considered harmful," *Communications of the ACM*, vol. 11, no. 3, pp. 147–148, http://www.acm.org/classics/oct95/, 1968.

[78] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins, "Self-similarity in the Web," *Proceedings of the International Conference on Very Large Data Bases*, http://citeseer.ist.psu.edu/dill01selfsimilarity.html, 2001.

[79] A. Dix, A. Friday, B. Koleva, T. Rodden, H. Muller, C. Randell, and A. Steed, "Managing multiple spaces," in *Spaces, Spatiality and Technology*, (P. Turner and E. Davenport, eds.), Dordrecht: Kluwer, 2005. http://www.equator.ac.uk/index.php/articles/c94/.

[80] D. Donato, L. Laura, S. Leonardi, and S. Millozzi, "Large scale properties of the webgraph," *European Physical Journal B*, vol. 38, pp. 239–243, 2004.

[81] L. Downes and C. Mui, *Unleashing the Killer App: Digital Strategies for Market Dominance*, Cambridge MA: Harvard Business School Press, 2000.

[82] H. L. Dreyfus and S. E. Dreyfus, "Making a mind versus modelling the brain: Artificial intelligence back at a branch-point," *Artificial Intelligence*, vol. 117, no. 1, 1988.

[83] B. Du, M. Demmer, and E. Brewer, "Analysis of WWW traffic in Cambodia and Ghana," *Proceedings of WWW 2006*, http://www2006.org/programme/files/pdf/5510.pdf, 2006.

[84] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins, "Visualizing tags over time," *Proceedings of WWW 2006*, http://www2006.org/programme/files/pdf/25.pdf, 2006.

[85] W. H. Dutton and A. Shepherd, "Confidence and risk on the Internet," in *Trust and Crime in Information Societies*, (R. Mansell and B. S. Collins, eds.), pp. 207–244, Cheltenham: Edward Elgar, 2005.

[86] eBay, "Evaluating a Member's Reputation," http://pages.ebay.com/help/feedback/evaluating-feedback.html, 2005.

[87] S. A. Edwards, "It's alive," *Wired*, Apr.97 1997.

[88] J. Ellman, "Corporate ontologies as information interfaces," *IEEE Intelligent Systems*, pp. 79–80, Jan/Feb 2004.

[89] P. Erdös and A. Rényi, "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Science*, vol. 5, pp. 17–61, 1960.

[90] P. Evans and T. S. Wurster, *Blown to Bits: How the New Economics of Information Transforms Strategy*, Cambridge MA: Harvard Business School Press, 2000.

[91] R. Fagin, R. Kumar, K. S. McCurley, J. Novak, D. Sivakumar, J. A. Tomlin, and D. P. Williamson, "Searching the workplace Web," *Proceedings of the 12th International World Wide Web Conference*, http://www2003.org/cdrom/papers/refereed/p641/xhtml/p641-mccurley.html, 2003.

[92] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," *ACM SIGCOMM 99*, vol. 29, 1999.

[93] J. Farrell and C. Shapiro, "Intellectual property, competition and information technology," in *The Economics of Information Technology: An Introduction*, (H. R. Varian, J. Farrell, and C. Shapiro, eds.), pp. 49–86, Cambridge: Cambridge University Press, 2004.

[94] M. Fayed, P. Krapivsky, J. Byers, M. Crovella, D. Finkel, and S. Redner, "On the emergence of highly variable distributions in the autonomous system topology," *ACM Computer Communication Review*, July 2003.

[95] M. Fayzullin, M. Nanni, D. Pedraschi, and V. S. Subrahmanian, "Foundations of distributed interaction systems," *Annals of Mathematics and Artificial Intelligence*, vol. 28, pp. 127–168, http://citeseer.ist.psu.edu/478943.html, 2000.

[96] D. Fensel, *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, 2nd edition*, Berlin: Springer, 2004.

[97] D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, "Introduction," in *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential*, (D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, eds.), pp. 1–25, Cambridge MA: MIT Press, 2003.

[98] A. Ferrara, "Web services: A process algebra approach," in *Proceedings of the 2nd International Conference on Service-Oriented Computing (ICSOC 2004)*, (M. Aiello, M. Aoyama, F. Curbera, and M. P. Papazoglou, eds.), pp. 242–251, New York: ACM Press, 2004.

[99] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, "A large-scale study of the evolution of Web pages," *Software: Practice and Experience*, vol. 34, no. 2, pp. 213–237, http://research.microsoft.com/research/sv/sv-pubs/p97-fetterly/p97-fetterly.html, 2004.

[100] J. H. Fetzer, "Program verification: The very idea," *Communications of the ACM*, vol. 31, pp. 1048–1063, 1988.

[101] G. W. Flake, D. M. Pennock, and D. C. Fain, "The self-organized Web: The yin to the Semantic Web's yang," *IEEE Intelligent Systems*, vol. 18, no. 4, http://research.yahoo.com/publication/OR-2003-003.pdf, 2003.

[102] I. Foster and C. Kesselman, eds., *The Grid 2: Blueprint for a New Computing Infrastructure*, San Francisco: Morgan Kaufmann, 2003.

[103] S. Fox, J. Q. Anderson, and L. Rainie, "The Future of the Internet: In a Survey, Technology Experts and Scholars Evaluate Where the Network is

Headed in the Next Ten Years," Pew Internet & American Life Project, http://www.pewinternet.org/pdfs/PIP_Future_of_Internet.pdf, 2005.

[104] C. Fry, M. Plusch, and H. Lieberman, "Static and dynamic semantics of the Web," in *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential*, (D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, eds.), pp. 377–401, Cambridge MA: MIT Press, 2003.

[105] X. Fu, T. Bultan, and J. Su, "Analysis of interacting BPEL Web services," *Proceedings of the World Wide Web Conference 2004*, 2004.

[106] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider, "Sweetening ontologies with DOLCE," in *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, (A. Gómez-Pérez and V. R. Benjamins, eds.), pp. 166–181, Berlin: Springer, 2002.

[107] J. Garrett and D. Waters, "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information," The Commission on Preservation and Access, and the Research Libraries Group, http://www.rlg.org/ArchTF/, 1996.

[108] W. Gaver, A. Boucher, S. Pennington, and B. Walker, "Evaluating technologies for ludic engagement," *CHI '05 Workshop on Affective Evaluation*, http://www.equator.ac.uk/index.php/articles/c94/, 2005.

[109] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cognitive Science*, vol. 7, no. 2, pp. 155–170, 1983.

[110] Y. Gil, "Knowledge mobility: Semantics for the Web as a white knight for knowledge-based systems," in *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential*, (D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, eds.), pp. 253–278, Cambridge MA: MIT Press, 2003.

[111] Y. Gil and V. Ratnakar, "Trusting information sources one citizen at a time," *Proceedings of the 1st International Semantic Web Conference (ISWC)*, 2002.

[112] A. Ginsberg and D. Hirtle, eds., *RIF Use Cases and Requirements*, 2006. http://www.w3.org/TR/rif-ucr/.

[113] J. A. Goguen, "Ontology, ontotheology and society," *International Conference on Formal Ontology in Information Systems (FOIS 2004*, http://charlotte.ucsd.edu/users/goguen/pps/fois04.pdf, 2004.

[114] J. Golbeck and B. Parsia, "Trust network based filtering of aggregated claims," *International Journal of Metadata, Semantics and Ontologies*, vol. 1, no. 1, http://trust.mindswap.org/papers/ijmso.pdf, 2005.

[115] J. Golbeck, B. Parsia, and J. Hendler, "Trust networks on the Semantic Web," in *Proceedings of the 7th International Workshop on Cooperative Intelligent Agents*, (M. Klusch, S. Ossowski, A. Omicini, and H. Laamenen, eds.), pp. 238–249, Berlin: Springer-Verlag, 2003. http://www.mindswap.org/papers/CIA03.pdf.

[116] S. Grabner-Kräuter and E. A. Kaluscha, "Empirical research in on-line trust: A review and critical assessment," *International Journal of Human-Computer Studies*, vol. 58, pp. 783–812, 2003.

[117] P. Graham, "A Plan for Spam," http://www.paulgraham.com/spam.html, 2002.

[118] T. L. Griffiths and J. B. Tenenbaum, "Optimal predications in everyday cognition," *Psychological Science*, http://web.mit.edu/cocosci/Papers/prediction10.pdf, 2006.

[119] G. Grimmett, *Percolation*, Berlin: Springer, 2nd edition ed., 1989.

[120] G. Grimnes, P. Edwards, and A. Preece, "Learning from semantic flora and fauna," *Proceedings of the AAAI Workshop on Semantic Web Personalization*, http://maya.cs.depaul.edu/∼mobasher/swp04/accepted/grimnes.pdf, 2004.

[121] W. I. Grosky, D. V. Sreenath, and F. Fotouhi, "Emergent semantics and the multimedia Semantic Web," *ACM Sigmod*, vol. 31, no. 4, pp. 54–58, http://lsdis. cs.uga.edu/SemNSF/SIGMOD-Record-Dec02/Gorsky.pdf (sic), 2002.

[122] P. Groth, S. Miles, V. Tan, and L. Moreau, eds., *Architecture for Provenance Systems Version 0.4*. 2005. http://eprints.ecs.soton.ac.uk/11310/.

[123] T. Gruber, "A translation approach to formal ontologies," *Knowledge Acquisition*, vol. 5, no. 25, pp. 199–200, http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html, 1993.

[124] N. Guarino, "Formal ontology, conceptual analysis and knowledge representation," in *Formal Ontology in Conceptual Analysis and Knowledge Representation: Special Issue of International Journal of Human-Computer Studies*, (N. Guarino and R. Poli, eds.), 1995. http://citeseer.ist.psu.edu/guarino95formal.html.

[125] N. Guarino and C. A. Welty, "An overview of OntoClean," in *The Handbook on Ontologies*, (S. Staab and R. Studer, eds.), pp. 151–172, Berlin: Springer-Verlag, 2004.

[126] P. Haase, M. Ehrig, A. Hotho, and B. Schnizler, "Personalized information access in a bibliographic peer-to-peer system," *Proceedings of the AAAI Workshop on Semantic Web Personalization*, http://maya.cs.depaul.edu/∼mobasher/swp04/accepted/haase.pdf, 2004.

[127] W. Hall, "The button strikes back," *New Review of Hypermedia and Multimedia*, vol. 6, pp. 5–17, 2000.

[128] T. Hammond, T. Hamay, B. Lund, and J. Scott, "Social bookmarking tools (I): A general review," *D-Lib*, vol. 11, no. 4, http://www.dlib.org/dlib/april05/hammond/04hammond.html, 2005.

[129] S. Harnad, "The symbol grounding problem," *Physica D*, vol. 42, pp. 335–346, http://cogprints.org/3106/01/sgproblem1.html, 1990.

[130] S. Harnad, *The symbol grounding problem*, Nature Publishing Group/Macmillan, 2003. http://cogprints.org/3018/01/symgro.htm.

[131] S. Harnad, "Publish or perish – self-archive to flourish: The green route to open access," *ERCIM News*, http://eprints.ecs.soton.ac.uk/11715/, 2006.

[132] T. Hart and L. Fazzani, *Intellectual Property Law 3rd Edition*, Basingstoke: Palgrave Macmillan, 2004.

[133] J. Haugeland, "Understanding natural language," *Journal of Philosophy*, vol. 76, pp. 619–632, 1979.

[134] J. A. Hendler, "Frequently Asked Questions on W3C's Web Ontology Language (OWL)," http://www.w3.org/2003/08/owlfaq, 2004.

[135] J. A. Hendler, "From Atoms to Owls: The New Ecology of the WWW," Keynote Lecture, XML2005, http://www.cs.umd.edu/∼hendler/presentations/XML2005Keynote.pdf, 2005.

[136] N. Henze and M. Kriesell, "Personalization functionality for the Semantic Web: Architectural outline and first sample implementations," *Proceedings of 1st International Workshop on Engineering the Adaptive Web*, http://rewerse.net/publications/download/REWERSE-RP-2004-31.pdf, 2004.

[137] M. Henzinger and S. Lawrence, "Extracting knowledge from the World Wide Web," *PNAS*, vol. 101, http://www.pnas.org/cgi/reprint/101/suppl_1/5186, April 6th 2004.

[138] M. R. Henzinger, "Algorithmic challenges in Web search engines," *Internet Mathematics*, vol. 1, no. 1, pp. 115–126, 2004.

[139] M. R. Henzinger, R. Motwani, and C. Silverstein, "Challenges in Web search engines," *SIGIR Forum*, http://www.sigir.org/forum/F2002/henzinger.pdf, Fall 2002.

[140] C. Hewitt, *PLANNER: A Language for Manipulating Models and Proving Theorems in a Robot*, AI Memo AIM-168, MIT, 1970. http://hdl.handle.net/1721.1/6171.

[141] C. A. R. Hoare, *Communicating Sequential Processes*, New York: Prentice-Hall, 1984.

[142] D. L. Hoffman, T. P. Novak, and A. Venkatesh, "Has the Internet become indispensable?," *Communications of the ACM*, vol. 47, no. 7, pp. 37–42, 2004.

[143] D. L. Hoffman, T. P. Novak, and A. Venkatesh, "Has the Internet Become Indispensable? Empirical Findings and Model Development," Working Paper, Sloan Center for Internet Retailing, Vanderbilt University, http://elab.vanderbilt.edu/research_papers.htm, 2004.

[144] M. Hollis, *Trust Within Reason*, Cambridge: Cambridge University Press, 1998.

[145] J. B. Horrigan, "Online Communities: Networks that Nurture Long-Distance Relationships and Social Ties," Pew Internet and American Life Project, http://www.pewinternet.org/pdfs/PIP_Communities_Report.pdf, 2001.

[146] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen, "From SHIQ and RDF to OWL: The making of a Web ontology language," *Journal of Web Semantics*, vol. 1, no. 1, http://www.websemanticsjournal.org/volume1/issue1/Horrocksetal2003/index.html, 2003.

[147] M. Horstmann, M. Lorenz, A. Watkowski, G. Ioannidis, O. Herzog, A. King, D. G. Evans, C. Hagen, C. Schlieder, A.-M. Burn, N. King, H. Petrie, S. Dijkstra, and D. Crombie, "Automated interpretation and accessible presentation of technical diagrams for blind people," *New Review of Hypermedia and Multimedia*, vol. 10, no. 2, pp. 141–163, 2004.

[148] J. Huang and M. S. Fox, "Uncertainty in knowledge provenance," *Proceedings of 1st European Semantic Web Symposium*, http://www.eil.utoronto.ca/km/papers/EuroSemWeb04-online.pdf, 2004.

[149] Z. Huang and H. Stuckenschmidt, "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, 2005.

[150] B. A. Huberman and L. A. Adamic, "Growth dynamics of the World-Wide Web," *Nature*, vol. 401, p. 131, 1999.

[151] B. A. Huberman and L. A. Adamic, "Information dynamics in the networked world," in *Complex Networks*, (E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, eds.), pp. 371–398, Berlin: Springer, 2003. http://www.hpl.hp.com/research/idl/papers/infodynamics/infodynamics.pdf.

[152] D. Huynh, N. R. Jennings, and N. R. Shadbolt, "Developing an integrated trust and reputation model for open multi-agent systems," *Proceedings of the 7th International Workshop on Trust in Agent Societies*, http://eprints.ecs.soton.ac.uk/9557/01/aamas-trust04.pdf, 2004.

[153] J. Iria and F. Ciravegna, "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, 2005.

[154] I. Jacobs, ed., *Technical Architecture Group (TAG) Charter*. 2004. http://www.w3.org/2004/10/27-tag-charter.

[155] I. Jacobs and N. Walsh, eds., *Architecture of the World Wide Web Volume One*. 2004. http://www.w3.org/TR/webarch/.

[156] A. Jaimes, "Human factors in automatic image retrieval design and evaluation," *SPIE Conference: Electronic Imaging 2006*, http://www.ee.columbia.edu/∼ajaimes/Pubs/ajaimes_spie06.pdf, 2006.

[157] X. Jin, Y. Zhou, and B. Mobasher, "A unified approach to personalization based on probabilistic latent semantic models of Web usage and content," *Proceedings of the AAAI Workshop on Semantic Web Personalization*, http://maya.cs.depaul.edu/∼mobasher/swp04/accepted/jin.pdf, 2004.

[158] K. S. Jones, "What's new about the Semantic Web? some questions," *SIGIR Forum*, vol. 38, no. 2, http://www.acm.org/sigir/forum/2004D/sparck_jones_sigirforum_2004d.pdf, 2004.

[159] A. Jøsang, "A logic for uncertain probabilities," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 3, pp. 279–311, http://security.dstc.edu.au/papers/logunprob.pdf, 2001.

[160] A. Jøsang and D. McAnally, "Multiplication and comultiplication of beliefs," *International Journal of Approximate Reasoning*, vol. 38, no. 1, pp. 19–51, http://security.dstc.edu.au/papers/JM2004-IJAR.pdf, 2004.

[161] L. Kagal, T. Finin, M. Paolucci, N. Srinivasan, K. Sycara, and G. Denker, "Authorization and privacy for Semantic Web services," *IEEE Intelligent Systems*, pp. 52–58, July/August 2004.

[162] D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, 1982.

[163] D. R. Karger and D. Quan, "What would it mean to blog on the Semantic Web," *Journal of Web Semantics*, vol. 3, no. 2, http://www.websemanticsjournal.org/ps/pub/2005-18, 2005.

[164] N. Kavantzas, D. Burdett, G. Ritzinger, T. Fletcher, Y. Lafon, and C. Barreto, "Web Services Choreography Description Language Version 1.0," http://www.w3.org/TR/2005/CR-ws-cdl-10-20051109/, 2005.

[165] J. Kay and A. Lum, "Ontology-based user modelling for the Semantic Web," *Proceedings of the Workshop on Personalisation on the Semantic Web: PerSWeb '05*, pp. 15–23, http://www.win.tue.nl/persweb/full-proceedings.pdf, 2005.

[166] O. Kharif, "Less impact from the "Slashdot effect"," *BusinessWeek Online*, http://www.businessweek.com/technology/content/mar2005/tc2005032_0932_tc119.htm?campaign_id=search, 2nd Mar 2005.

[167] A. Kilgarrif and G. Grefenstette, "Introduction to the special issue on the Web as corpus," *Computational Linguistics*, vol. 29, no. 3, pp. 333–348, http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WAC Intro.pdf, 2003.

[168] J. Kim, "Supervenience and nomological incommensurables," *American Philosophical Quarterly*, vol. 15, pp. 149–156, 1978.

[169] J. Kim, "Psychological supervenience," *Philosophical Studies*, vol. 41, pp. 51–70, 1982.

[170] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing and retrieval," *Journal of Web Semantics*, vol. 2, no. 1, http://www.websemanticsjournal.org/ps/pub/2005-10, 2005.

[171] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 668–677, http://www.cs.cornell.edu/home/kleinber/auth.pdf, 1998.

[172] G. Klyne and J. J. Carroll, eds., *Resource Description Framework (RDF): Concepts and Abstract Syntax*, 2004. http://www.w3.org/TR/rdf-concepts/.

[173] G. F. Knolmayer and T. Myrach, "Concepts of Bitemporal Database Theory and the Evolution of Web Documents," *Institute of Information Systems, University of Bern, Working Paper 127*, http://www.ie.iwi.unibe.ch/publikationen/berichte/resource/WP-127.pdf, 2000.

[174] K. R. Koedinger and J. R. Anderson, "Abstract planning and perceptual chunks: Elements of expertise in geometry," *Cognitive Science*, vol. 14, pp. 511–550, 1990.

[175] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for emerging cyber-communities," *Proceedings of the 8th World Wide Web Conference*, http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html, 1999.

[176] A. M. Lai, J. Nieh, B. Bohra, V. Nandikonda, A. P. Surana, and S. Varshneya, "Improving Web browsing on wireless PDAs using thin-client computing," *Proceedings of World Wide Web Conference 2004*, http://www2004.org/proceedings/docs/1p143.pdf, 2004.

[177] C. Lam, "Collaborative filtering using associative neural memory," *Proceedings of the AAAI Workshop on Semantic Web Personalization*, http://maya.cs.depaul.edu/∼mobasher/swp04/accepted/lam.pdf, 2004.

[178] A. N. Langville and C. D. Meyer, "Deeper inside PageRank," *Internet Mathematics*, vol. 1, no. 3, http://www.internetmathematics.org/volumes/1/3/Langville.pdf, 2004.

[179] O. Lassila and M. Adler, "Semantic gadgets: Ubiquitous computing meets the Semantic Web," in *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential*, (D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, eds.), pp. 363–376, Cambridge MA: MIT Press, 2003.

[180] D. Lee, H. Chuang, and K. Seamons, "Effectiveness of document ranking and relevance feedback techniques," *IEEE Software*, vol. 14, no. 2, pp. 67–75, 1997.

[181] R. Lempel and S. Moran, "The stochastic approach for link structure analysis (SALSA) and the TKC effect," *Proceedings of the 9th World Wide Web Conference*, 2000.

[182] R. Lempel and A. Soffer, "PicASHOW: Pictorial authority search by hyperlinks on the Web," *Proceedings of the 10th World Wide Web Conference*, 2001.

[183] D. B. Lenat, "Cyc: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, 1995.

[184] A. Lenhart, "Around 25 Million People Have Used the Internet to Sell Something," Pew Internet and American Life Project, http://www.pewinternet.org/pdfs/PIP_SellingOnline_Nov05.pdf, 2005.

[185] E. L. Lesser, M. A. Fontaine, and J. A. Slusher, eds., *Knowledge and Communities*, Boston: Butterworth-Heinemann, 2000.

[186] L. Lessig, *Code and Other Laws of Cyberspace*, New York: Basic Books, 1999.

[187] L. Lessig, *The Future of Ideas: The Fate of the Commons in a Connected World*, New York: Random House, 2001.

[188] S.-T. A. Leung, S. E. Perl, R. Stata, and J. L. Wiener, "Towards Web-Scale Web Archaeology," Compaq Systems Research Center report #174, 2001.

[189] J. Liang, R. Kumar, and K. W. Ross, "Understanding KaZaA," Working paper, http://cis.poly.edu/∼ross/papers/UnderstandingKaZaA.pdf, 2004.

[190] J. Lohse, K. Biolsi, N. Walker, and H. Rueter, "A classification of visual representations," *Communications of the ACM*, vol. 37, no. 12, pp. 36–49, 1994.

[191] A. López-Ortiz, "Algorithmic foundations of the Internet," *ACM SIGACT News*, vol. 36, no. 2, 2005.

[192] D. MacKenzie, *Mechanizing Proof: Computing, Risk and Trust*, Cambridge MA: MIT Press, 2001.

[193] T. Maekawa, T. Hara, and S. Nishio, "Image classification for mobile Web browsing," *Proceedings of WWW 2006*, http://www2006.org/programme/files/pdf/2506.pdf, 2006.

[194] T. Maneewatthana, G. Wills, and W. Hall, "Adaptive personal information environment based on Semantic Web," *Proceedings of the International Workshop on Adaptive and Personalized Semantic Web, Hypertext 2005*, http://www.ru5.cti.gr/HT05/files/Maneewatthana.pdf, 2005.

[195] F. Manola and E. Miller, eds., *RDF Primer*, 2004. http://www.w3.org/TR/rdf-primer/.

[196] C. C. Marshall and F. L. Shipman, "Which Semantic Web?," in *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, pp. 57–66, ACM Press, 2003. http://www.csdl.tamu.edu/∼marshall/ht03-sw-4.pdf.

[197] D. K. Mauzy and R. S. Milne, *Singapore Politics Under the People's Action Party*, London: Routledge, 2002.

[198] D. L. McGuinness and F. van Harmelen, eds., *OWL Web Ontology Language Overview*, 2004. http://www.w3.org/TR/owl-features/.

[199] S. Mendes and R. P. Chaves, "Enriching WordNet with qualia information," *Proceedings of NAACL 2001 Workshop on WordNet and Other Lexical Resources*, http://errepe.no.sapo.pt/ewqf.pdf, 2001.

[200] P. Mika, "Flink: Semantic Web technology for the extraction and analysis of social networks," *Journal of Web Semantics*, vol. 3, no. 2, http://www.websemanticsjournal.org/ps/pub/2005-20, 2005.

[201] P. Mika, "Ontologies are us: A unified model of social networks and Semantics," *Proceedings of the 4th International Semantic Web Conference (ISWC2005)*, 2005.

[202] T. Miles-Board, L. Carr, and W. Hall, "Looking for linking: Associative links on the Web," *Proceedings of 13th ACM Conference on Hypertext and Hypermedia (HT '02)*, http://eprints.ecs.soton.ac.uk/6977/, 2002.

[203] R. Milner, *Communication and Concurrency*, New York: Prentice Hall, 1989.

[204] R. Milner, *Communicating and Mobile Systems: The Pi-Calculus*, Cambridge: Cambridge University Press, 1999.

[205] B. A. Misztal, *Trust in Modern Societies*, Cambridge: Polity Press, 1996.

[206] H. Müller, P. Clough, W. Hersh, T. Deselaers, T. M. Lehmann, B. Janvier, and A. Geissbuhler, "Using heterogeneous annotation and visual information for the benchmarking of image retrieval systems," *SPIE Conference Photonics West: Electronic Imaging*, http://medir.ohsu.edu/∼hersh/spie-06-imageclef.pdf, 2006.

[207] J. Myers, "What can the Semantic Grid do for science and engineering?," *Proceedings of Dagstuhl Seminar on Semantic Grid: The Convergence of Technologies*, http://drops.dagstuhl.de/opus/volltexte/2005/395/pdf/05271.MyersJames.ExtAbstract.395.pdf, 2005.

[208] S. Narayanan and S. McIlraith, "Analysis and simulation of Web services," *Computer Networks*, vol. 42, no. 5, pp. 675–693, 2003.

[209] D. Neef, G. A. Siesfeld, and J. Cefola, eds., *The Economic Impact of Knowledge*, Boston: Butterworth-Heinemann, 1998.

[210] F. Neuhaus, P. Grenon, and B. Smith, "A formal theory of substances, qualities and universals," in *Formal Ontology in Information Systems*, (A. Varzi and L. Vieu, eds.), pp. 49–59, Turin: IOS Press, 2004.

[211] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," http://protege.stanford.edu /publications/ontology_development/ontology101-noy-mcguinness.html, 2001.

[212] N. F. Noy and M. A. Musen, "The PROMPT suite: Interactive tools for ontology merging and mapping," *International Journal of Human-Computer Studies*, vol. 59, no. 6, pp. 983–1024, 2003.

[213] N. F. Noy, W. Grosso, and M. A. Musen, "Knowledge-acquisition interfaces for domain experts: An empirical evaluation of Protégé-2000," *12th International Conference on Software Engineering and Knowledge Engineering (SEKE2000)*, http://smi-web.stanford.edu/auslese/smi-web/reports/SMI-2000-0825.pdf, 2000.

[214] C. Nugent, D. Doyle, and P. Cunningham, "Gaining Insight Through Case-Based Explanation," Technical Report TCD-CS-2004-49, Trinity College Dublin, https://www.cs.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-49.pdf, 2004.

[215] K. O'Hara, *Trust: From Socrates to Spin*, Cambridge: Icon Books, 2004.

[216] K. O'Hara, H. Alani, and N. Shadbolt, "Identifying communities of practice: Analysing ontologies as networks to support community recognition," in *Proceedings of the 2002 World Computer Congress, Information Systems: The E-Business Challenge*, (R. Traunmüller, ed.), pp. 89–102, Dordrecht: Kluwer, 2002. http://eprints.ecs.soton.ac.uk/6522/.

[217] K. O'Hara and D. Stevens, *inequality.com: Power, Poverty and the Digital Divide*, Oxford: OneWorld, 2006.

[218] E. T. O'Neill, B. F. Lavoie, and R. Bennett, "Trends in the evolution of the public Web 1998–2002," *D-Lib Magazine*, vol. 9, no. 4, http://www.dlib.org/dlib/april03/lavoie/04lavoie.html, 2003.

[219] E. T. O'Neill, P. D. McClain, and B. F. Lavoie, "A Methodology for Sampling the World Wide Web," Online Computer Library Center, http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003447, 1998.

[220] A. Oram, *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, Sebastopol, CA: O'Reilly & Associates, 2001.

[221] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Dept. of Computer Science, Stanford University, technical report 1999-66, 1999.

[222] T. S. Parikh and E. D. Lazowska, "Designing an architecture for delivering mobile information services to the rural developing world," *Proceedings of WWW 2006*, http://www2006.org/programme/files/pdf/5507.pdf, 2006.

[223] J. Parsons, P. Ralph, and K. Gallagher, "Using viewing time to infer user preference in recommender systems," *Proceedings of the AAAI Workshop on Semantic Web Personalization*, http://maya.cs.depaul.edu/~mobasher/swp04/accepted/parsons.pdf, 2004.

[224] K. Pastra and Y. Wilks, "Vision-language integration in AI: A reality check," *Proceedings of ECAI 2004*, http://www.dcs.shef.ac.uk/~yorick/papers/ecai04.pdf, 2004.

[225] J. Pathak, D. Caragea, and V. G. Honovar, "Ontology-extended component-based workflows: A framework for constructing complex workflows from semantics heterogeneous software components," *Proceedings of the International Workshop on Semantic Web and Databases (SWDB-04)*, http://www.cs.iastate.edu/~honavar/Papers/SWDB04.pdf, 2004.

[226] W. D. Perdue, "The new totalitarianism: Cyber-hegemony and the global system," *International Roundtable on the Challenges of Globalization*, http://-p-o.org/perdue.htm, 1999.

[227] M. Perkowitz and O. Etzioni, "Towards adaptive Web sites: Conceptual framework and case study," *Proceedings of the 8th World Wide Web Conference*, 1999.

[228] E. L. Peterson, "Customized resource discovery: Linking formalized Web taxonomies to a Web ontology hub," *Proceedings of the AAAI Workshop on Semantic Web Personalization*, http://maya.cs.depaul.edu/∼mobasher/swp04/accepted/peterson.pdf, 2004.

[229] K. Petridis, F. Precioso, T. Athanasiadis, Y. Avrithis, and Y. Kompatsiaris, "Combined domain specific and multimedia ontologies for image understanding," *Proceedings of the Workshop on Mixed-Reality as a Challenge to Image Understanding and Artificial Intelligence, 28th German Conference on Artificial Intelligence*, http://www.image.ntua.gr/papers/384.pdf, 2005.

[230] F. Piper, M. J. B. Robshaw, and S. Schwiderski-Grosche, "Identities and authentication," in *Trust and Crime in Information Societies*, (R. Mansell and B. S. Collins, eds.), pp. 91–112, Cheltenham: Edward Elgar, 2005.

[231] G. Priest, "Paraconsistent logic," in *Handbook of Philosophical Logic 2nd Edition Vol.6*, (D. F. Gabbay and F. Guenthner, eds.), pp. 287–393, Dordrecht: Kluwer Academic Publishers, 2002.

[232] E. Prud'hommeaux and A. Seaborne, eds., *SPARQL Query Language for RDF*, 2005. http://www.w3.org/TR/rdf-sparql-query/.

[233] H. Putnam, "The meaning of "meaning"," in *Mind, Language and Reality: Philosophical Papers Volume 2*, pp. 215–271, Cambridge: Cambridge University Press, 1975.

[234] C. D. Raab, "The future of privacy protection," in *Trust and Crime in Information Societies*, (R. Mansell and B. S. Collins, eds.), pp. 282–318, Cheltenham: Edward Elgar, 2005.

[235] D. Rafiei and A. Mendelzon, "What is this page known for? computing web page reputations," *Proceedings of the 9th World Wide Web Conference*, 2000.

[236] L. Rainie, "Big Jump in Search Engine Use," Pew Internet and American Life Project, http://www.pewinternet.org/pdfs/PIP_SearchData_1105.pdf, 2005.

[237] L. Rainie, "Use of Web Cams," Pew Internet and American Life Project, http://www.pewinternet.org/pdfs/PIP_webcam_use.pdf, 2005.

[238] L. Rainie and M. Madden, "Podcasting Catches on," Pew Internet & American Life Project, http://www.pewinternet.org/pdfs/PIP_podcasting.pdf, 2005.

[239] S. D. Ramchurn and N. R. Jennings, "Trust in agent-based software," in *Trust and Crime in Information Societies*, (R. Mansell and B. S. Collins, eds.), pp. 165–204, Cheltenham: Edward Elgar, 2005.

[240] R. Reiter, "Equality and domain closure in first-order databases," *Journal of the ACM*, vol. 27, no. 2, pp. 235–249, 1980.

[241] D. Resnick, "Politics on the internet: The normalization of cyberspace," in *The Politics of Cyberspace*, (C. Toulouse and T. W. Luke, eds.), pp. 48–68, New York: Routledge, 1998.

[242] P. Resnick and R. Zeckhauser, "Trust among strangers in Internet transactions: Empirical analysis of e-Bay's reputation system," in *The Economics of the Internet and E-Commerce: Advances in Applied Microeconomics Vol.11*, (M. R. Baye, ed.), pp. 127–157, Amsterdam: Elsevier Science, 2002.

[243] M. Richardson, R. Agrawal, and P. Domingos, "Trust management for the Semantic Web," *Proceedings of the 2nd International Semantic Web Conference*, http://www.cs.washington.edu/homes/pedrod/papers/iswc03.pdf, 2003.

[244] P. Rodriguez, S. Mukherjee, and S. Rangarajan, "Session level techniques for improving Web browsing performance on wireless links," *Proceedings of the World Wide Web Conference 2005*, http://www2004.org/proceedings/docs/1p121.pdf, 2004.

[245] O. Roy, *Globalized Islam: The Search for a New Ummah*, New York: Columbia University Press, 2004.

[246] J. Sabater and C. Sierra, "REGRET: A reputation model for gregarious societies," in *Proceedings of 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*, (C. Castelfranchi and L. Johnson, eds.), pp. 475–482, New York: ACM Press, 2002.

[247] P. Samuelson and S. Scotchmer, "The law and economics of reverse engineering," *Yale Law Journal*, vol. 111, pp. 1575–1663, http://socrates.berkeley.edu/∼scotch/re.pdf, 2002.

[248] E. Sanchez, ed., *Fuzzy Logic and the Semantic Web*, Amsterdam: Elsevier, 2006.

[249] S. Saroiu, K. P. Gummadi, R. Dunn, S. D. Gribble, and H. M. Levy, "An analysis of internet content delivery systems," *Proceedings of the 5th Symposium on Operating Systems Design and Implementation*, 2002.

[250] C. Schindler, P. Arya, A. Rath, and W. Slany, "htmlButler – wrapper usability enhancement through ontology sharing and large scale cooperation," *Proceedings of the International Workshop on Adaptive and Personalized Semantic Web, Hypertext 2005*, http://www.ru5.cti.gr/HT05/files/rath.pdf, 2005.

[251] M. M. C. Schraefel, N. R. Shadbolt, N. Gibbins, H. Glaser, and S. Harris, "CS AKTive Space: Representing computer science on the Semantic Web," *Proceedings of WWW 2004*, http://www.www2004.org/proceedings/docs/1p384.pdf, 2004.

[252] R. Schroeder, A. Caldas, G. Mesch, and W. Dutton, "The World Wide Web of Science: Reconfiguring access to information," *Proceedings of the Annual Conference of the National Centre for e-Social Science*, http://www.oii.ox.ac.uk/research/files/W3Sc_ncess2005_paper_Schroeder.pdf, 2005.

[253] N. Schwartz, R. Cohen, D. ben-Avraham, A.-L. Barabási, and S. Havlin, "Percolation in directed scale-free networks," *Physics Review E*, vol. 66, http://www.wisdom.weizmann.ac.il/∼recohen/publications/directed.pdf, 2002.

[254] P. Seabright, *The company of strangers: A natural history of economic life*, Princeton: Princeton University Press, 2004.

[255] J. Searle, "Minds, brains and programs," *The Behavioral and Brain Sciences*, vol. 3, pp. 417–424, 1980.

[256] J. Seidenberg and A. Rector, "Web ontology segmentation: Analysis, classification and use," *Proceedings of WWW 2006*, http://www2006.org/programme/files/pdf/4026.pdf, 2006.

[257] N. Shadbolt, W. Hall, and T. Berners-Lee, "The Semantic Web revisited," *IEEE Intelligent Systems*, pp. 96–101, May/June 2006.

[258] G. Shafer, "Causal logic," *Proceedings of IJCAI-98*, http://www.glennshafer.com/assets/downloads/article62.pdf, 1998.

[259] C. Shirky, "Ontology is Overrated: Categories, Links and Tags," http://www.shirky.com/writings/ontology_overrated.html, 2005.

[260] N. Simou, C. Saathoff, S. Dasiopoulou, E. Spyrou, N. Voisine, V. Tzouvaras, I. Kompatsiaris, Y. Avrithis, and S. Staab, "An ontology infrastructure for multimedia reasoning," *International Workshop on Very Low Bit-Rate Video-Coding*, http://www.image.ntua.gr/papers/381.pdf, 2005.

[261] B. Skyrms, *Evolution of the Social Contract*, Cambridge: Cambridge University Press, 1996.

[262] J. Slaney, "Relevant logic and paraconsistency," in *Inconsistency Tolerance*, (L. Bertossi, A. Hunter, and T. Schaub, eds.), pp. 270–293, Berlin: Springer, 2004.

[263] A. Sloman, "Diagrams in the mind?," in *Diagrammatic Representation and Reasoning*, (M. Anderson, B. Meyer, and P. Olivier, eds.), London: Springer-Verlag, 2001. http://www.cs.bham.ac.uk/research/cogaff/sloman.diagbook.pdf.

[264] A. Smeaton, "Creating information links in digital video as a means to support effective video navigation," *keynoteMultimedia Information Retrieval Workshop, SIGIR 2003*, http://km.doc.ic.ac.uk/mmir2003/1alansmeaton wsmm03.pdf, 2003.

[265] B. Smith, "Protecting consumers and the marketplace: The need for federal privacy legislation," speech to the Congressional Internet Caucus, http://www.netcaucus.org/speakers/2005/smith/privacyspeech.pdf, 2005.

[266] J. F. Sowa and A. K. Majumdar, "Analogical reasoning," in *Conceptual Structures for Knowledge Creation and Communication*, (A. de Moor, W. Lex, and B. Ganter, eds.), Berlin: Springer, 2003. http://www.jfsowa.com/pubs/analog.htm.

[267] J. Stanley and T. Williamson, "Knowing how," *Journal of Philosophy*, vol. 98, pp. 411–444, 2001.

[268] L. Steels, "Semiotic dynamics for embodied agents," *IEEE Intelligent Systems*, pp. 32–38, May/June 2006.

[269] H. Störrle, *Models of Software Architecture: Design and Analysis With UML and Petri Nets*, Ph.D. thesis, University of München, 2000.

[270] K. Su, A. Sattar, G. Governatori, and Q. Chen, "A computationally grounded logic of knowledge, belief and certainty," *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2005)*, http://eprint.uq.edu.au/archive/00002328/01/p409-su.pdf, 2005.

[271] V. S. Subrahmanian, "Amalgamating knowledge bases," *ACM Transactions on Database Systems*, vol. 19, no. 2, pp. 291–331, 1994.

[272] C. Sunstein, *republic.com*, Princeton: Princeton University Press, 2001.

[273] J. B. H. Tan and J. S. L. Yong, "Many agencies, one government – Singapore's vision of public services delivery," in *E-Government in Asia: Enabling Public Service Innovation in the 21st Century*, (J. S. L. Yong, ed.), pp. 267–308, Singapore: Marshall Cavendish Business, 2003.

[274] P. Twomey, "Hey governments, hands off the internet," *New Scientist*, 12th Nov 2005.

[275] A. Uszok, J. M. Bradshaw, M. Johnson, R. Jeffers, A. Tate, J. Dalton, and S. Aitken, "Kaos policy management for Semantic Web services," *IEEE Intelligent Systems*, pp. 32–41, http://www.aiai.ed.ac.uk/project/ix/documents/2004/2004-ieee-is-uszok-kaos.pdf, July/August 2004.

[276] W. M. P. van der Aalst, "Pi calculus versus petri nets: Let us eat "humble pie" rather than further inflate the "pi hype"," *BPTrends*, vol. 3, no. 5, pp. 1–11, http://is.tm.tue.nl/staff/wvdaalst/pi-hype.pdf, 2005.

[277] P. C. van Fenema and F. Go, "Dispersed communities: Weaving collective and individual life, fact and fiction, mediated and direct spatial experiences," *Communities and Technologies Conference*, http://www.fbk.eur.nl/PEOPLE/pfenema/personal/, 2003.

[278] J. van Ossenbruggen, L. Hardman, and L. Rutledge, "Hypermedia and the Semantic Web: A research agenda," *Journal of Digital Information*, vol. 3, no. 1, http://jodi.ecs.soton.ac.uk/Articles/v03/i01/VanOssenbruggen/, 2002.

[279] J. van Ossenbruggen, G. Stamou, and J. Z. Pan, "Multimedia annotations and the Semantic Web," *Workshop on Semantic Web Case Studies and Best Practice for eBusiness, International Semantic Web Conference*, http://homepages.cwi.nl/∼media/publications/SWCASE05.pdf, 2005.

[280] C. J. van Rijsbergen, "Evaluation," in *Information Retrieval 2nd Edition*, 1979. online book, http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html.

[281] H. R. Varian, "Competition and market power," in *The Economics of Information Technology: An Introduction*, (H. R. Varian, J. Farrell, and C. Shapiro, eds.), pp. 1–47, Cambridge: Cambridge University Press, 2004.

[282] J. M. Vidal, P. Buhler, and C. Stahl, "Multiagent systems with workflows," *IEEE Internet Computing*, pp. 76–82, http://jmvidal.cse.sc.edu/papers/vidal04a.pdf, January/February 2004.

[283] R. Volz, S. Handschuh, S. Staab, L. Stojanovic, and N. Stojanovic, "Unveiling the idden bride: Deep annotation for mapping and migrating legacy data to the Semantic Web," *Journal of Web Semantics*, vol. 1, no. 2, http://www.websemanticsjournal.org/ps/pub/2004-15, 2004.

[284] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small world" networks," *Nature*, vol. 393, pp. 440–442, 1998.

[285] G. Weimann, *Terror on the Internet: The New Arena, The New Challenges*, Washington: United States Institute of Peace, 2006.

[286] D. J. Weitzner, "Testimony Before the United States Senate Committee on Commerce, Science and Transportation," http://www.w3.org/2000/05/25-Senate-Privacy-Testimony.html, 2000.

[287] D. J. Weitzner, J. Hendler, T. Berners-Lee, and D. Connolly, "Creating a Policy-Aware Web: Discretionary, rule-based access for the World Wide Web," in *Web and Information Security*, (E. Ferrari and B. Thuraisingham, eds.), Hershey PA: Idea Group Inc, 2005. http://www.mindswap.org/users/hendler/2004/PAW.html.

[288] WGIG, "Report of the Working Group on Internet Governance," http://www.wgig.org/docs/WGIGREPORT.pdf (available in a number of languages and formats from http://www.wgig.org/, 2005.

[289] Y. Wilks, "Ontotherapy, or how to stop worrying about what there is," in *Knowledge Representation With Ontologies: Present Challenges, Future Possibilities*, (C. Brewster and K. O'Hara, eds.), International Journal of Human-Computer Studies, 2006.

[290] Y. Wilks, "The Semantic Web as the apotheosis of annotation, but what are its semantics?," in press. http://www.dcs.shef.ac.uk/∼yorick/papers/AAAI_Paper.pdf.

[291] L. Wittgenstein, *Philosophical Investigations*, Oxford: Basil Blackwell, 1953.

[292] L. Wittgenstein, *Remarks on the Foundations of Mathematics 3rd Edition*, Oxford: Basil Blackwell, 1978.

[293] K. Yanai and K. Barnard, "Probabilistic Web image gathering," *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR 2005)*, http://kobus.ca/research/publications/ACM-MM-MIR-05/Yanai-Barnard-ACM-MM-MIR-05.pdf, 2005.

[294] B. Yuwono and D. Lee, "Search and ranking algorithms for locating resources on the World Wide Web," *Proceedings of the 12th International Conference on Data Engineering*, 1996.

[295] L. Zadeh, "From search engines to question-answering systems – the problems of world knowledge, relevance, deduction and precisiation," *Keynote2005 IEEE International Conference on Information Reuse and Integration*, http://www.cs.fiu.edu/IRI05/, 2005.

[296] J. Zhang, J.-Y. Chung, C. K. Chang, and S. W. Kim, "WS-Net: A Petri-net based specification model for Web services," Proceedings of IEEE International Conference on Web Services (ICWS '04), 2004.

[297] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic models for discovering e-communities," *Proceedings of WWW 2006*, http://www2006.org/programme/files/pdf/1016.pdf, 2006.

[298] M. Zhou and K. Venkatesh, *Modeling, Simulation and Control of Flexible Manufacturing Systems: A Petri Net Approach*, Singapore: World Scientific Publishing, 1999.

[299] H. Zhuge, L. Zheng, N. Zhang, and X. Li, "An automatic Semantic relationships discovery approach," *poster presentationWorld Wide Web Conference 2004*, http://www2004.org/proceedings/docs/2p278.pdf, 2004.