Web Science

Kieron O’Hara and Wendy Hall

Version of a chapter to appear in William Dutton (ed.), *The Oxford Handbook of Internet Studies*, Oxford: Oxford University Press, 2012

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

Abstract: This chapter examines some of the ideas behind the emerging discipline of Web Science, whose ambition is to shape the future development of the World Wide Web and the research agenda that that requires. There are formidable obstacles to this ambition, not least the large scale of the Web, the most complex piece of technology ever devised, and the co-constitution of the Web with its communities of users. Web Science must therefore straddle and integrate computing, mathematics, complexity and network studies on the one hand, together with studies of the social context, using the methods of sociology, law and economics on the other. The chapter defines the Web and differentiates it from the Internet, upon whose infrastructure it depends. Web Science is shown to be a type of reflective practice, made problematic by its scale and by the complexity of its interrelation with embedding societies; in particular it is difficult to integrate the micro-scale of the protocols which define it, with the macro-scale of the social effects that follow from widespread use of particular systems. The example of the development of the Web of Linked Data is used to illustrate difficulties and potential solutions

Introduction: the rationale for Web Science

The World Wide Web is an extraordinarily transformative technology. Claims for its significance range from hype to scepticism, but most agree that its capacity for supporting communication and access to documents is orders of magnitude beyond previous technologies, bringing great changes not only to the Internet and ICT, but also to the offline world, affecting the media, entertainment, politics and government, science and research, administration and commerce. Whole new areas of activity such as social networking or e-crime have flourished using its protocols. The number of users is vast and growing, and its decentralised structure – there is no editor of content, no quality control and anyone can link to anything – has democratised communication in all sorts of ways.

Yet for all that the Web is remarkably under-studied and under-theorised. There seem to be three principal reasons for this. First, it is a dauntingly large and complex structure. Understanding the Web in its context requires working at a number of scales from the micro-level of the detail of individual protocols like HTTP (HyperText Transfer Protocol) or HTML (HyperText Markup Language), to the macro-level of emergent behaviour such as blogging, spamming or e-commerce. Second, it evolves very quickly, so data soon become outdated. Third, it is a curious amalgam of technologies (hardware, software protocols, and programming environments such as Java and AJAX [Asynchronous JavaScript And XML]) and human activities (the Web links not only documents and data, but people as well).

Hence a comprehensive overview demands multi-disciplinary skills relevant to computing, law, economics, sociology, management and organisation studies, media studies, semiotics, mathematics, and innumerable sub-disciplines (Berners-Lee et al 2006a). Too often the Web is studied as an example of a particular phenomenon – a network, a set of computer languages, or a platform for commerce. It is all those, but taken as a whole it is so much more.

In particular, we must not fall into the error of thinking that the proper study of the Web is within computer science (CS). In CS, Web-related research focuses on technical issues such as information retrieval algorithms or the algorithms for routing information through the underlying Internet. Yet these properties, important as they are, cannot be the whole story. Google's PageRank link analysis algorithm, for instance, is a brilliant piece of work, but its significance to the Web depends not only on the algorithmic structure, but also on the *context* of its use, which is outside the province of CS. Nothing about the algorithm *per se* explains how the eigenvectors that it computes map miraculously onto the conversations that Web users have, nor about how it can constantly be adapted to avoid spoofing.

Many research memes within CS are positively hostile to the Web's governing principles. For example, consider the famous letter written by the formalist Edsger Dijkstra entitled 'GOTO Statement Considered Harmful' (Dijkstra 1968). In that, he argued (correctly) that the sudden and arbitrary leaps that the GOTO command made possible rendered the formalisation of programs extremely difficult, and therefore the use of the command should be avoided in programming for critical systems. Yet, in effect, hyperlinks mean the Web is *constituted* by the wretched GOTOs!

How is the Web likely to develop? Sensemaking, reuse and retrieval are vital. It is currently largely made up of linked documents, often text documents, so Natural Language Processing techniques add value by extracting some form of meaning from the human-readable text of the pages based on heuristics or statistics (cf. Wilks & Brewster 2009). But an increasingly important extension, the *Semantic Web*, envisages linking *data* resources enriched by ontologies which give interpretations of terms used, to allow machine processing of the Web's content (Shadbolt et al 2006). This development is exciting yet challenging. How can we allow independent consistent data systems to be connected locally without requiring an implausible and totalitarian *global* consistency? How do we query an unbounded Web of linked information repositories? How should we align different data models, and visualise and navigate the huge connected graph of information that results? Who should bear liability for shared data resources?

To answer such questions and understand the basic issues underlying them, researchers are fostering the new discipline of *Web Science* (Berners-Lee et al 2006a, 2006b, Shneiderman 2007, Shadbolt & Berners-Lee 2008) to develop methods and curricula to understand the Web and provide foundations for engineering methodologies. Web Science is not just modelling the Web. It includes engineering infrastructure protocols using tools from many disciplines which may involve radical thinking about technology and society, but must respect the Web's invariants: decentralisation to avoid bottlenecks and facilitate scalability; serendipitous reuse of information; fairness, openness and trust (Berners-Lee et al 2006a). If Web Science delivers a greater understanding of the Web, threats can be identified and addressed, opportunities pursued, and the Web itself can be adapted to social change.

This chapter explores the agenda of Web Science for the development of the future Web. We begin with a definition of the object of study, the architecture and conventions of the World Wide Web itself. The next section explores the foundational assumptions of Web Science, looking in the abstract at how the Web's development can be influenced. The following section takes as an example the role of Web Science in the development of a Web of Linked Data, before rounding off with a concluding discussion.

What exactly *is* the Web?

The distinction between the Internet and the Web is not widely understood; the Web is certainly the most visible Internet application, and most Internet users are also Web users, so the two are often confused. In this section, we will briefly set out the simple technologies which make the Web a flexible, usable information space which, most importantly, scales when the number of users increases (see Jacobs & Walsh 2004 for more detail). Our task in this section is therefore to set out the essential technologies and protocols that make up the Web, and the social regularities that have helped it flourish.

Web architecture

The Web is a space in which *resources* are identified by *Uniform Resource Identifiers* (URIs – Berners-Lee et al 2005). *Protocols* support *interaction* and *information transfer* between computers, while *formats* are used to *represent* the information resources. These are the basic ingredients of the Web, upon whose designs depends the utility and efficiency of Web interaction.

Identification of resources is essential for sharing information about them, reasoning about them, modifying or exchanging them. Resources may be anything that can be linked to or spoken of. Not all resources are on the Web. Even if they are *identifiable* from the Web, they 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 digital artefact (for example, a text file or a video) are called *information resources*.

For reasoning and referencing to happen on a global scale, an identification system is required to provide a single global standard; URIs provide that system. It would be possible to develop alternatives to URIs, but a *single* universal system of identifiers facilitates linking, bookmarking and other value-adding functions across heterogeneous applications. Ideally each URI identifies a single resource in a context-independent manner (this is desirable, but not enforceable). Accessing a resource via a URI is called *dereferencing* the URI.

URIs fall under particular defined *schemes*, of which the most commonly used are *HTTP*, *FTP* (File Transfer Protocol) and *mailto:*. If we take HTTP as an example, an HTTP URI should ideally refer to a single resource, and be allocated to a single owner. What accessing a resource entails varies from context to context, but a common experience is receiving a representation of the (state of the) resource on a browser. It need not be this way: it may be that no representation of the resource is available, or that access is limited (e.g. password controlled). Not all types of URI 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 development of the Web as a space for *interaction* follows from the ability of agents 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 identification and tracking of obligations incurred online through unsafe interactions.

The power of the Web in enabling communication, free expression, querying and interaction 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, thereby allowing 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 authors to particular URI schemes. Although a stable reference system will reduce ambiguities, allow consistent reference to resources of whatever type across heterogeneous applications and facilitate the automation of information-retrieval tasks, URI schemes cannot be enforced. For a name in a public language to be successful, it must be adopted by a community which has some tacit agreement on its use (Halpin 2009). It is not essential to have a well-ordered set of names–advances in statistically-based search techniques mean that much information can be retrieved relatively efficiently–but the value of the best current search techniques is proportional to the quality of links.

Finally, it is an essential principle of Web architecture that errors should be handled simply and flexibly. Errors are inevitable in an information space of thousands of terabytes, whose users number in the billions. If dangling links (URIs with no resource at the end of them), ill-formed content or other predictable errors caused the system to crash it would never have functioned in the first place. Furthermore, interoperability requires that agents should be able to recover from errors without compromising awareness that the error has occurred. Hence a dangling link, for example, will merely return the irritating but hardly fatal '404 error'.

Conventional aspects of Web use

It would be incorrect to see Web architecture as a core topic, and social and ethical questions as 'bolt ons' to be addressed after the fact. These latter are fundamental. The Web is a deliberately decentralised structure, which means that there is no authority to enforce good behaviour. Many types of behaviour essential for the Web to work (meaning, convention, commitment) are understandable from the point of view of rational self-interest (Skyrms 1996), but there are payoffs to bad behaviour, of commission (opportunities to gain by cheating) and omission (failure to maintain a website satisfactorily). Hence self-interested rationality cannot *entirely* explain how such cooperative behaviour gets off the ground (Hollis 1998; Seabright 2004); for the Web to exist demands *social norms* as well as technical protocols. These social norms have doubtless evolved partly in the context of previous mass telecommunications technologies, which will therefore have an indirect effect upon the development of the Web (cf. Perkins & Neumayer 2011).

Web Science can help determine what practices and conventions are essential, 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 'desirable' or even 'moral' online behaviour. Social norms and engineering turn out to be linked, and may have profound consequences for the Web's future (O'Hara 2009). Some have even suggested that value be embedded in design (Baken et al 2010), although the Web's decentralisation will make that hard, and probably undesirable, to enforce.

Let's consider the example of the connection between a URI and a resource in more detail. As anyone who has had to maintain a Website will know, pressure to change URIs builds up. One diachronic study of 150m webpages found that after 9 weeks access was lost to over 10% of them

(Fetterly et al 2004). Some degradation is caused by genuine engineering difficulties, some is merely sloth, but the Web will function better if URIs don't change, and always point to the same document (or its latest version).

Avoiding changing URIs is easier said than done. For instance, when a website is reorganised, the temptation is to provide a neat new rationalised structure, expressed as a new set of URIs, expressing the new organisational philosophy. It is tempting, but unwise, to create directories called 'latest' or 'current' which will become outdated. Dangling links are frustrating, and do a lot to undermine trust in websites and companies (Grabner-Kräuter & Kaluscha 2003). Any record of a URI by an interested party, whether a bookmark, a link from another site, or a scribbled note on paper, records the URI of a page at a moment in time, and cannot easily be automatically updated (Berners-Lee 1998).

Hence convention collides with Web engineering. One incurs obligations and duties when online because of the cooperative nature of the Web, and sustaining the Web's important invariants depends on them being taken seriously. Lessig (1999) has correctly argued that behaviour can be constrained online by architecture, regulation and market-based incentives, but he is careful to emphasise that social norms play a vital part as well. One of the goals of Web Science is to be able to provide models of behaviour and architecture that allow different types of constraint to be virtually explored and experimented with. We are a long way from achieving a unified view, but investigations of these problematic cases are important early steps.

The science of the Web

One common misconception about engineering is that it is the application of scientific theory to achieve desired ends, given prior agreement on framing a problem and on the ends. Yet because of the *sui generis* nature of many engineering problems–this certainly applies to the Web of course–much of the essential knowledge needed for a solution must be derived in practice, often in response to unforeseen challenges perceived during a project itself. This has led to the development of a theory of design and engineering called *reflective practice* (Schön 1983).

In this methodology, the problem as initially set is not fixed, as practitioners must change their perceptions and strategies in response to uncertainty, instability and unique features of the problem. They proceed experimentally, to create and discover new solutions that need be neither unique nor optimal. Controlled, reversible experiments are out of the question; the Web (as with other major engineering projects) is not a closed system and any large-scale intervention will tend to change the object of study itself. Hence each experiment that the engineer tries must be as far as possible sensitive to the needs of the context, and take into account understanding of the social and psychological context–theoretical knowledge about complex systems cannot be tested in isolated, closed subsystems.

Web Science as reflective practice

Web Science is a type of reflective practice (O’Hara & Hall 2010). Given the complexity of the problem space, it will be essential to develop engineering methods that use the insights of reflective practice, dynamically and recursively reconfiguring the problem specification as knowledge is gained during the design and engineering processes.

Engineering the Web requires sensitivity to both technical and social concerns. The designer has an idea for an innovation and develops protocols, formalisms, software and hardware to realise the vision, which may or may not be formally or precisely specified. However, no digital system lives in a vacuum, and its use will depend on a number of assumptions about social context implicit in the design. Note that the designer cannot specify every aspect of the system’s behaviour; at some point assumptions about context will have to carry some functional weight. For instance, the email system SMTP (simple mail transfer protocol) was developed on the basis of assumptions about what people would want communications to carry, about organisational context and about the motives of senders (specifically that messages would be sent in good faith by a homogeneous academic community all of whose members would be concerned with the same group of problems, so messages would be relevant to the receiver, generated in response to a genuine requirement, with a transparent meaning).

On the Web, however, unintended consequences at the macro-level can emerge from changes at the micro-level. For example, as more users take up a system, there might be a marked and noticeable change in social behaviour. Analysis of these macro-level effects is likely to uncover new social issues, which need to be addressed in their turn–and one way of doing this is to design and build new technology, leading to another cycle of design and social change.

To continue the example of SMTP, when it became a macro phenomenon used by people beyond the target community, the unintended consequences of free and simple communication became clear. Problems such as spam and phishing began to emerge. Social changes also accompanied the technology. Emails leave a semi-permanent record so it became harder for companies to hide their internal decision-making, scientists their suppression of data, and errant spouses their infidelities. New technical solutions, such as spam filtering, were now needed to solve the problems created by the emergent phenomena. These developments have been accompanied by parallel adjustments in the law, corporate best practice, and our intuitive understanding of privacy (McArthur 2001) which themselves raise more issues (cf. e.g. O’Hara & Shadbolt 2008), and so the cycle continues.

The dynamics and topography of the Web

The characterisation of Web engineering as a cyclical conversation between scientists and engineers, users and techies, fits neatly into Schön’s (1983) ideas about reflective practice. However, the position is not as simple as this makes it appear (cf. O’Hara & Hall 2010). Although the Web shares some of its developmental characteristics with other telecommunications technologies (Perkins & Neumayer 2011), the *singularities* of the Web as a piece of designed technology demand its intensive study as a first order object as envisaged by the Web Science programme.

Consider the zone of time in which an action may make a difference, what Schön calls the *action-present* (1983, 62), which depends on the pace of activity and the boundaries of potential action. For the Web, this is both tiny and vast, depending on point of view. The cycles of Web development are measured in years. Blogging, for instance, took a number of years to develop from small beginnings, and then 'suddenly' took off at the beginning of the century. 'Suddenly' in this case takes us from the appearance of the first blogging tools and guides and the first major political issues influenced by bloggers in 2001 and 2002, to the exponential growth characteristic of the years after 2004. But we also need to factor in the timescale of an effective intervention. What seems imperative in year 0 of a research project may be completely out of date by year 3 when a product appears.

New types of online behaviour can become very popular very quickly. At the time of writing, Facebook and Twitter dominate thinking about cutting-edge large-scale Web phenomena (cf. e.g. Gaffney 2010), but by, say, 2015 it is quite likely that the landscape will be very different and those giants will appear hopelessly out of date. As each new star application comes along, new users (who may have been children during the previous cycle) will arrive with it, rendering older assumptions void. In short, what might seem a relatively long action-present for Web Science is in reality very attenuated. By the time data are gathered, models created and simulations run, the opportunity to influence events may already be past.

Hence Web Science must be concerned not only with topography but also the dynamics of the Web. There are a number of technologies and methods for mapping the Web (see Thelwall, this volume). What do such maps tell us (cf. e.g. Donato et al 2004)? The visualisations are often 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? Furthermore, the Web is not a static information space, but rather is dynamic and evolving (O'Neill et al 2003), and models should ideally have built into them 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 perpetually prone to revision (cf. e.g. Barabási et al 2000).

The rapid growth of the Web made a complete survey out of the question years ago. In such circumstances, representative sampling is important, but how should a sample be gathered in order to be properly called representative (Leung et al 2001)? To be properly useful, a sample should be *random*; but what does ‘random’ mean here? Are we concerned, for instance, with websites or webpages? Furthermore, so cheap are operations on the Web that a small number of operators can skew results however carefully the sample is chosen. One survey (Fetterly et al 2004) discovered that 27% of pages in Germany’s .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 users in Germany, but rather that over a million URIs, most but not all on servers registered in the German domain, 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 facie* 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. Kilgarrif & Grefenstette 2003), and how does increasing cleverness in rendering affect things (Henzinger 2004)? Even if we were happy with our sampling methodology, how amidst all the noise could we discover interesting structures efficiently (López-Ortiz 2005)?

Web Science needs to take into account the variance of scale between intervention and outcome. Any experimental change will be relatively small scale – a new type of software, a new type of communications protocol, a small social network. The consequences of the change *relative to the intention of the innovation* can be described and studied in small-scale experiments in the lab, or with a small set of pioneer users. Such intentions are usually focused on the experience of a single type of user or organisation. The problem, of course, is that few if any of the *global* consequences of Web technologies are of this tractable type, because they affect very large groups of people and organisations, so that most consequences at the scale of the Web are unintended. What experiment could have predicted the phenomenal growth of Facebook? Early experiments among small social groups for particular purposes gave Mark Zuckerberg and colleagues early impetus, although "Thefacebook's Palo Alto geeks [including Zuckerberg] lacked confidence in their own judgments about how people would respond to the product" (Kirkpatrick 2010: 64). The geeks were shrewd; there was clearly no empirical basis for saying that (a) Facebook would have 600 million active users, (b) it would outperform apparently stronger rivals such as MySpace, (c) it would challenge and even change very basic social norms and concepts such as privacy and friendship, or (d) that Zuckerberg would be able to wield political influence with politicians such as Barack Obama and David Cameron seeking airtime with him?

The development of the Web and the role of Web Science: semantics and linked data

Berners-Lee's original Semantic Web vision argued that there is too little machine-readable information on the WWW as it was then constituted.

> *The meaning of the documents is clear to those with a grasp of (normally) English, and the significance of the links is only evident from the context around the anchor. To a computer, [on the other hand], the Web is a flat, boring world devoid of meaning. This is a pity, as in fact documents on the Web describe real objects and imaginary concepts, and give particular relationships between them. … Adding semantics to the Web involves two things: allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values. Only when we have this extra level of semantics will we be*

> *able to use computer power to help us exploit the information to a greater extent than our own reading.* (Berners-Lee 1994)

This vision of automation and machine-processability came to be dubbed the *Semantic Web* but an important preliminary stage, the *Linked Data Web*, is the release of linked and linkable data (Bizer et al 2009). Early adopters of linked data include e-science and e-social science, which depend on the integration and automatic interrogation of large quantities of distributed data (O'Hara et al 2010). Governments, such as the UK government in its data.gov.uk programme, have also shown an interest in the Linked Data Web as the medium for representing and releasing public data (Koumenides et al 2010, Shadbolt et al 2011). In this section, we will discuss the potential of the Linked Data Web, and the role of Web Science in facilitating it.

How does the Linked Data Web work?

The Linked Data Web relies on a series of formalisms and technologies. URIs provide a global naming convention for resources, as described above. The *Resource Description Framework* (RDF – Manola & Miller 2004) is a knowledge representation language that was designed with the Semantic Web in mind. Its basic format is a simple subject-predicate-object structure ("Brian is the child of Albert"), and because it has three elements an RDF statement is therefore called a *triple*. RDF assigns URIs to the subjects, predicates and objects that it links, allowing representation of data in such a way that anything referred to in the data (whether an object or a relation) can be linked to. Ideally, dereferencing the URIs should provide access to useful information about the resources, as well as useful links to other data.

Links can be made using various mechanisms, the simplest of which is a URI that points to another. For example (taken from Berners-Lee 2006/2009), someone might describe some relationships (that Albert is the father of Brian and Carol) in RDF as follows:

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

This RDF is about three resources which have local identifiers ‘#albert’, ‘#brian’ and ‘#carol’, and might be obtained from a file called ‘<http://example.org/smith>’. HTTP can be used to generate a globally-invariant identifier for the three resources; for instance “http://example.org/smith#albert” refers to #albert as defined in the named file, and so on. Now there is a global identifier, links can be made by anyone without ambiguity. For instance, a document ‘<http://example.org/jones>’ might contain the following RDF:

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

Here a series of relationships between resources #denise, #edwin and #carol have been asserted, but the datum about #carol links it to the data in the other file. Someone following the link can dereference the URI by decomposing ‘http://example.org/smith#carol’ into two parts: the part before the ‘#’ which gives the name and location of the file; and ‘#carol’ which is the local identifier in that file. Hence the information about #carol in the first file can be accessed thanks to the link included in the second file. This is the simplest way of linking data, though of course there are more complex methods (Berners-Lee 2006/2009).

In 2010, the Linked Open Data project counted 13 billion triples of linked data on the Web (Möller et al 2010). The ability to move between data linked in such a way opens up the possibility of exposing data on the Web and being able to access it from any application. The advantage of this is that when data from other sources is accessed, following the links gives the information user access to a contextualisation of the data, or to more information that can be exploited about the subject. If the data retrieved is also linked, then following *those* links gives access to more information, and so on. Linking data therefore allows the creation of an extremely rich context for an inquiry which furthermore can be interrogated automatically.

The value of linked data

The Web of Linked Data will change our model of the value of information. Currently, the value of information stems from its *scarcity*–people and organisations gain value from information they have gathered, and exploit monopoly rights via legal contrivances such as copyright, intellectual property rights, licensing, and so on. Even when organisations do not resort to the law, they make great investments in protecting trade secrets. However, this scarcity-based model seems inadequate for the digital age.

In the first place, the social benefits from unlicensed use of 'protected' knowledge and innovation, were already large in the pre-digital economy, and indeed account for much of our wealth today: "some 80 percent of the benefits [from innovation] may plausibly have gone to persons who made no direct contribution to innovation. The rather startling implication of all this is that the spillovers of innovation, both direct and indirect, can be estimated to constitute well over half of current [US] GDP–and it can even be argued that this is a very conservative figure" (Baumol 2002, 135). And secondly, the Web has made it harder to preserve monopoly rights to information as copying and distribution reduces the marginal cost to copiers to close to zero. Although many media companies have taken rearguard action to protect their IP, so simple is the distribution model on the Web that the basis of the value of information is rapidly switching from scarcity to *abundance*. It is the large quantity of data that can be placed in novel and unintended contexts with little cost that gives it value in the age of digital technologies–and the Linked Data Web is designed to foster such abundance.

Trust

The technical means to support the linking of data are necessary but insufficient for Web-scale adoption to realise the potential value. Many social mechanisms are required, including incentives for individuals and legal frameworks and protections, but *trust* is perhaps key to the spread of linked data. If information is to be drawn routinely from heterogeneous sources, then it is important that users are able to trust it in order to be able to act on the wider set of inferences they can make. Trust, which mediates risk, will depend on the criticality of the inferences and the risk-aversion of the trustor (O'Hara et al 2004, Bonatti et al 2006, Creese & Lambert 2009). Measuring trust, however, is a complex problem (Golbeck & Hendler 2004). An important parameter is the provenance of data (including statements about the methods of production and the organisation that carried them out). Methods are appearing to describe provenance in open systems (Moreau 2010), but more needs to be discovered about how information spreads across the Web, and therefore how it can be tracked and understood (Berners-Lee et al 2006a).

Online trust in general

Ideally trust and trustworthiness would be linked causally so that all and only trustworthy people/systems/data are trusted. This presents us with another set of Web Science research challenges (O'Hara & Hall 2008). How can we maintain the causal link using Web technology? What incentives and economic models are available to promote trust and trustworthiness together?

Offline and online trust have somewhat different properties, with con-men and masqueraders able to exploit different properties of the interactive context to undermine interlocutors' expectations. Online, the user labours under two important disadvantages. First, he or she is deprived of the complexity of signal available in the offline world. Online, the signals are basically the visual ones specified by the HTML source file of the page, augmented possibly by the roles played by the parties in the transaction (e.g. the website is that of one's bank). However, role-based trust is not a very secure foundation, as people often fail to verify roles (Dhamija et al 2006). Second, the designer of the website is in total control of the signals that it gives out; the user has little or no opportunity to engage the website in 'conversation', to see how it 'performs', to 'size it up', as we do offline when we are judging people.

This presents us with a second set of research challenges. How should trust be represented, maintained and repaired on the Web? What variables are important? Will these change as we move from human to artificial agents? What sort of institutions and methods will help online trust? Can information from social networks inject some objectivity (Szomszor et al 2007, Breslin et al 2009, Victor et al 2009)?

The social dynamic of online trust is an area requiring far more research, but one review focused on three *perceptual* factors that were particularly relevant. *Perception of credibility* is to do with honesty, expertise, predictability and reputation. *Ease of use* relates to the simplicity and design of the website. *Risk* is the perceived likelihood of an undesirable outcome (Corritore et al 2003). The first two factors in particular are strongly connected to the gathering and evaluation of signals of trustworthiness. This confirms the findings of an earlier study which found six major features that encouraged trust in e-commerce sites – the site's brand, seals of approval, ease of navigation, a fulfilling ordering experience, the site's presentation and the technologies used to create the website–again strongly connected with the signalling systems characteristic of local trust (Cheskin Research 1999; and see Connolly, this volume).

However, Web users are not particularly efficient at picking up the right signals that provide the causal connection between trust and trustworthiness. Dhamija and colleagues (2006) investigated the reasons why bogus sites work, and discovered that existing anti-phishing browser cues are ineffective. A participant group in that experiment made mistakes 40% of the time (even though primed to look out for phishing sites), and surprisingly neither age, gender nor computing experience were significant variables. The study showed that people are unaware of the sorts of signalling systems that have been developed to ensure trustworthiness (e.g. the padlock symbol to show that the page was delivered securely by SSL), or of the typical strategies of counterfeiters (e.g. using images to mask underlying text, or placing an SSL-padlock in the body of a webpage). Furthermore, users often failed to notice the *lack* of expected signals of trustworthiness. Attention to the needs of actual Web users leads to a further set of Web Science research challenges. How can secure systems be made usable and effective for consumers, given the limited knowledge and bounded rationality of Web users? Indeed, as Halpin et al (2010) argue, this is a vital question, given the increasingly strong bonds between extended human cognition and online information representation, where people outsource much of their model of the world and their memory to digital resources.

Trusting data

Trusting data requires understanding the way it was created, and the principles underlying its representational format. Information about these issues can be associated with data by *annotation* with *metadata*. Metadata are descriptive data about data, including 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.

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 a variety of suitably-annotated resources), archiving guidance (Cedars 2002), and identifying information (such as a unique reference number). Perhaps the most important use is to promote interoperability, allowing the combination of heterogeneous resources across platforms without loss of content. Schemata 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 these.

As to what metadata are required, much depends on the reasons for annotation and the demands of data users. For many purposes–for example, sharing digital photos–the metadata can be curated by volunteer communities, as the success of Web 2.0 sites like Flickr shows (Breslin et al 2009). More generally, interesting possibilities for metadata include time-stamping, provenance, uncertainty and licensing restrictions.

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 help in such judgments though need not necessarily resolve them. Representing confidence in reliability has always been difficult in epistemic logic. Approaches include: subjective logic, which represents an opinion as a real-valued triple (belief, disbelief, uncertainty) where the three items add up to 1 (Jøsang 2001, Jøsang & McAnally 2004, Ceolin et al 2010); grading based on qualitative judgements, although such qualitative grades can be given numerical interpretations and then reasoned about mathematically (Gil & Ratnakar 2002, Golbeck et al 2003); fuzzy logic (cf. Sanchez 2006); and probability (Huang & Fox 2004).

There are two main problems with annotation on the Web. The first is the difficulty of reasoning with metadata; the formalisms listed in the previous paragraph exhibit the common trade-off that the most expressive are the most difficult to use. Second, the task of annotating legacy data is an enormous, if not a Sisyphean, one. It has been argued that annotating the Web will require large-scale automatic methods, which will in turn require strong knowledge modelling commitments (Kiryakov et al 2005); 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, or automatically annotating legacy resources such as databases underlying the deep Web (Volz et al 2004).

The role of governments, and the political effects of the Web

Politics will inevitably loom large in Web Science, for a number of reasons. Firstly, although governments were generally somewhat slow in adopting the Web as a tool for communication and administration (Accenture 2004, Homburg 2008, Nixon 2010 sum up developments and challenges), they have been taking the lead in the population of the Web of data, particularly linked data. Various trends in governance have promoted this development (cf.Dunleavy et al 2006, Hood & Margetts 2007). In the United States, the need for transparency in the 2009 stimulus of the economy led to the creation of the data.gov site to host open government data. In the United Kingdom, a more ideological drive toward transparency was adopted by the government of Gordon Brown, and has been accelerated by the Coalition government of 2010, in order to improve the efficiency and accountability of public services, as well as to facilitate the use of information by activists and entrepreneurs (O’Hara 2011 is the most complete write-up of the UK government’s programme at the time of writing). The data.gov.uk site now hosts thousands of government datasets, many of which are available under the Open Government Licence (which is very non-prescriptive and modelled on Creative Commons). These third-generation transparency initiatives (see Fung et al 2007, 169, who places this in the context of history of transparency government) have been prominent in the push towards linked data, partly by the publication of data in linkable form, and partly by the efforts of a developer community to convert government open data to linked data (cf. Dickinson 2010). Meanwhile, in the wider EU, attempts have been made to implement Semantic Web technologies in e-government (Vitvar et al 2010), but there has been a shift towards an open data agenda here too at the time of writing.

The effects of these initiatives, which have a momentum of their own, have yet to be fully felt, and may not have been completely anticipated by governments. Nevertheless, they have been vital in increasing the amount of linked data on the Web. However, it is fair to say that there is an important role for Web Science in the risk analysis and management of potentially sweeping changes in governmental models to ones more appropriate to the digital era (Dunleavy et al 2006).

Secondly, it is already becoming clear that new means of communication can be used by people to circumvent official channels of information, and to organise and spread messages beyond narrow local circles; open data and transparency are part of that, but not the whole story. These tendencies of modern ICT were already evident when the Falun Gong began using email to organise itself in China, and text messaging helped coordinate protesters against President Estrada of the Philippines. Web 2.0 methods of communication have proved important in challenging entrenched governments, and some early commentaries on the Arab Spring series of revolutions that spread in 2011 have drawn attention to the protesters' use of Twitter (at the time of writing, there has been relatively little academic work on this, and some scepticism about how much microblogging had helped, as opposed merely to drawing the attention of the Western world to the protests – cf. Papic & Noonan 2011). Certainly the early use of microblogging in the protests against the re-election of President Ahmadinejad of Iran in 2009 turned out to be counterproductive. The government was able to paint the tech-savvy protesters as an unrepresentative elite. Furthermore, once the rest of the world cottoned onto the situation in Tehran, the protesters' Tweets were drowned out by the sheer volume of supportive Tweets from the United States and elsewhere, thereby neutralising the effect of the use of microblogging (Economist 2009).

And thirdly, to paraphrase a recent argument by Evgeny Morozov, although the Web has traditionally been seen as a tool for spreading liberalism and conversation (Berners-Lee et al 2006a), and as a counterhegemonic medium (Warf & Grimes 1997), it may also provide a means of repression (Morozov 2011). Maybe that will influence its development too. The resolution of arguments about both the revolutionary and the repressive aspects of the Web will demand input from Web Science, which could act as an authoritative voice in a field currently driven from journalism and the blogosphere.

Conclusions

The Web is currently under-theorised, despite being the core of the world's information infrastructure. Most approaches see it as an instance of a particular type of structure, whether a mathematical network, a social network, a medium of communication, a set of computing languages and protocols, a medium of exchange, an ecology, an unpoliced domain for rugged individualists, an anarchist's paradise, a locus of cultural hegemony or even a Great Pulsating Brain At The Centre Of The Multiverse. No doubt most or all of these are valid in many ways, but one is reminded very much of the parable of the blind men creating mutually exclusive theories of an elephant's anatomy purely by touch. Without disparaging the work of investigators working within a single disciplinary perspective, it is the contention of this chapter that transcending these individual perspectives will yield important results.

The Web is not an exogenous entity. As Marx's 11th Thesis on Feuerbach has it, "philosophers have hitherto only interpreted the world in various ways: the point is to change it." Surprisingly, many have studied the Web without considering that they could influence its development. It is an engineered technology, and so can be altered. Conversely, many engineers have succeeded in changing the Web, but if those changes are uninformed by an understanding of the wider consequences, this creates the risk of causing harm either to the Web itself or wider society (as some have argued with respect to Google and Facebook).

This chapter has discussed the discipline of Web Science, investigating the World Wide Web as a first order object of study using a catholic variety of methods. It has argued that the hybrid analysis/engineering nature of Web Science allies it with the design/engineering methodology of reflective practice, although the variety of scales at which the Web can be studied makes it peculiarly problematic. Nevertheless, Web Scientists can play a part in developing the future Web. In this chapter, we looked at the idea of developing the Linked Data Web in some detail; other developments of interest, which could not be covered in this space, include the mobile Web, the dissemination of the Web in the developing world, the Policy Aware Web, the Web as a trusted, secure and private space, the Semantic Grid as an e-science tool, mechanisms for governance, standards development and the design of new standards. All these areas have been the focus of Web Science study, and it is the hope of the community that research will ultimately deliver a more socially-conscious and socially-sensitive World Wide Web.

It is of course an important challenge for Web Science that studying the Web as a first order object does not result in the neglect of its technical, social, economic and political context. Arguments about, say, net neutrality or walled gardens are highly ideologically-charged, but the acceptability or otherwise of innovation will evolve alongside social attitudes, the social demands made on the Web (e.g. the increased demand for video), and the capacity of the technical infrastructure. No Web Scientist can afford to remain ignorant of issues such as the growth of social networking sites as a means of managing users' identities, or the failure of American network operators to deliver sufficient bandwidth in one of the most mature Web domains. Neither can the Web's intimate contribution to human psychology be neglected either (cf. Halpin et al 2010, which uses the philosophy of extended cognition to argue that the Web helps constitute human minds; a striking conclusion, even if, like Turkle 2011 or Lanier 2010, one is appalled by the prospect). When the Web Scientist examines what a future would be like in which all data is 'stored in the cloud', he or she cannot ignore the carbon emissions of those data warehouses, with their insatiable appetite for air conditioning (in 2007, the carbon emissions of the ICT industry as a whole were actually equivalent to those of the airline industry, i.e. about 2% of the world's emissions–Climate Group 2008: 17).

The Web is an important system, highly contested between different interests, whose development can only be partially steered by invested authority, and which has been an object of ideological dispute almost as long as it has been in existence. The amalgamation of many disciplines is essential for understanding it (a) in full, and (b) in context. Ultimately, the aim of Web Science is to create, by education, training and practice, a research and engineering community within which diverse methods of analysis and synthesis are routinely integrated. Rather than a multidisciplinary approach to a single complex object, a measure of its success will be its acceptance as a discipline in its own right.

Acknowledgements

The work reported in this chapter was partly supported by the projects LiveMemories – Active Digital Memories of Collective Life, Bando Grandi Progetti 2006, Provincia Autonoma di Trento, and the EU FET project Living Knowledge, (http://livingknowledge-project.eu/), contract no. 231126. Some of the text has been derived from Tim Berners-Lee, Wendy Hall, James A. Hendler, Kieron O'Hara, Nigel Shadbolt & Daniel J. Weitzner, 'A framework for Web Science'.

References

Accenture (2004). *eGovernment Leadership: High Performance, Maximum Value*, Accenture.

Baken, N.H.G., Wiegel, V. & van Oortmerssen, G. (2010). 'The Value (Driven) Web'. In *2nd Web Science Conference*, Raleigh, N.C., http://journal.webscience.org/309/2/websci10_submission_50.pdf.

Barabási, A.-L., Albert, R. & Jeong, H. (2000). 'Scale-Free Characteristics of Random Networks: the Topology of the World Wide Web'. *Physica A* 281: 69-77.

Baumol, W.J. (2002). *The Free-Market Innovation Machine: Analyzing the Growth Miracle of Capitalism*, Princeton: Princeton University Press.

Berners-Lee, T. (1994). Plenary at WWW Geneva 94, http://www.w3.org/Talks/WWW94Tim/.

Berners-Lee, T. (1998). *Cool URIs Don't Change*. http://www.w3.org/Provider/Style/URI.

Berners-Lee, T. (2006/2009). *Linked Data*. http://www.w3.org/DesignIssues/LinkedData.html.

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

Berners-Lee, T., Hall, W., Hendler, J.A., O'Hara, K., Shadbolt, N. & Weitzner, D.J. (2006a). 'A Framework for Web Science'. *Foundations and Trends in Web Science*, 1/1: 1-130.

Berners-Lee, T., Hall, W., Hendler, J.A., Shadbolt, N. & Weitzner, D.J. (2006b). 'Creating a Science of the Web'. *Science*, 313/5788: 769-771.

Bizer, C., Heath, T. & Berners-Lee, T. (2009). 'Linked Data – the Story so Far'. *International Journal on Semantic Web and Information Systems*, 5/3: 1-22.

Bonatti, P.A., Duma, C., Fuchs, N., Nejdl, W., Olmedilla, D., Peer, J. & Shahmehri, N. (2006). 'Semantic Web Policies – a Discussion of Requirements and Research Issues'. In Y. Sure & J. Domingue (eds.), *The Semantic Web: research and applications, 3rd European Semantic Web Conference 2006 (ESWC-06)*, Springer: Berlin.

Breslin, J.G., Passant, A. & Decker, S. (2009). *The Social Semantic Web*, Berlin: Springer.

Cedars (2002). *Cedars Guide to Preservation Metadata*, http://www.leeds.ac.uk/cedars/guideto/metadata/guidetometadata.pdf.

Ceolin, D., van Hage, W.R. & Fokkink, Wan (2010). 'A Trust Model to Estimate the Quality of Annotations Using the Web'. In *2nd Web Science Conference*, Raleigh, N.C., http://journal.webscience.org/315/2/websci10_submission_81.pdf.

Cheskin Research and Studio Archetype/Sapient (1999). *eCommerce Trust Study*, http://www.cheskin.com/cms/files/i/articles//17__report-eComm%20Trust1999.pdf.

The Climate Group (2008). *Smart2020: Enabling the Low Carbon Economy in the Information Age*, http://www.theclimategroup.org/_assets/files/Smart2020Report.pdf.

Corritore, C.L., Kracher, B. & Weidenbeck, S. (2003). 'On-line Trust: Concepts, Evolving Themes, a Model'. *International Journal of Human-Computer Studies*, 58: 737-758.

Creese, S. & Lamberts, K. (2009). 'Can Cognitive Science Help Us Make Information Risk More Tangible Online?' In *1st Web Science Conference*, Athens, Greece, http://journal.webscience.org/142/2/tangibility_of_risk_websci09_FINAL.pdf.

Dhamija, R., Tygar, J.D. & Hearst, M. (2006). 'Why Phishing Works'. In *Conference on Human Factors in Computing Systems (CHI 2006)*, http://people.seas.harvard.edu/~rachna/papers/why_phishing_works.pdf.

Dickinson, I. (2010). 'COINS as Linked Data'. http://data.gov.uk/resources/coins.

Dijkstra, E. (1968). 'Go To Statement Considered Harmful'. *Communications of the ACM*, 11/3: 147-148.

Donato, D., Laura, L., Leonardi, S. & Millozzi, S. (2004). 'Large Scale Properties of the Webgraph'. *European Physical Journal B* 38: 239-243.

Dunleavy, P., Margetts, H., Bastow, S. & Tinkler, J. (2006). *Digital Era Governance: IT Corporations, the State, and E-Government*, Oxford: Oxford University Press.

The Economist (2009). 'Twitter 1, CNN 0'. *The Economist*, 18th June, 2009.

Fetterly, D., Manasse, M., Najork, M. & Wiener, J. (2004). 'A Large-Scale Study of the Evolution of Web Pages'. *Software: Practice and Experience*, 34/2: 213-237, http://research.microsoft.com/research/sv/sv-pubs/p97-fetterly/p97-fetterly.html.

Fung, A., Graham, M. & Weil, D. (2007). *Full Disclosure: The Perils and Promise of Transparency*, New York: Cambridge University Press.

Gaffney, D. (2010). '#iranElection: Quantifying Online Activism'. In *2nd Web Science Conference*, Raleigh, N.C., http://journal.webscience.org/295/2/websci10_submission_6.pdf.

Gil, Y. & Ratnakar, V. (2002). 'Trusting Information Sources One Citizen at a Time'. In *Proceedings of the 1st International Semantic Web Conference (ISWC)*.

Golbeck, J. & Hendler, J (2004). 'Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks'. In E. Motta, N. Shadbolt, A. Stutt & N. Gibbins (eds.), *Engineering Knowledge in the Age of the Semantic Web, Proceedings of 14th International Conference, EKAW 2004*, Springer: Berlin: 116-131.

Golbeck, J., Parsia, B. & Hendler, J (2003). 'Trust Networks on the Semantic Web'. In M. Klusch, S. Ossowski, A. Omicini & H. Laamenen (eds.), *Proceedings of the 7th International Workshop on Cooperative Intelligent Agents*, Berlin: Springer-Verlag: 238-249, http://www.mindswap.org/papers/CIA03.pdf.

Grabner-Kräuter, S. & Kaluscha, E.A. (2003). 'Empirical Research in On-line Trust: a Review and Critical Assessment'. *International Journal of Human-Computer Studies*, 58: 783-812.

Halpin, H. (2009). 'Social Meaning on the Web: from Wittgenstein to Search Engines'. In *1st Web Science Conference*, Athens, Greece, http://journal.webscience.org/190/3/halpin-websci09.pdf.

Halpin, H., Clark, A. & Wheeler, M. (2010). 'Towards a Philosophy of the Web: Representation, Enaction, Collective Intelligence'. In *2nd Web Science Conference*, Raleigh, N.C., http://journal.webscience.org/324/.

Henzinger, M.R. (2004). 'Algorithmic Challenges in Web Search Engines'. *Internet Mathematics* 1/1: 115-126.

Hollis, M. (1998). *Trust Within Reason*, Cambridge: Cambridge University Press.

Homburg, V. (2008). *Understanding E-Government: Information Systems in Public Administration*, Abingdon: Routledge.

Hood, C.C. & Margetts, H.Z. (2007). *The Tools of Government in the Digital Age*, Basingstoke: Palgrave Macmillan.

Huang, J. & Fox, M.S. (2004). 'Uncertainty in Knowledge Provenance'. In *Proceedings of 1st European Semantic Web Symposium*, http://www.eil.utoronto.ca/km/papers/EuroSemWeb04-online.pdf.

Jacobs, I. & Walsh, N. (eds.) (2004). *Architecture of the World Wide Web Volume One*, http://www.w3.org/TR/webarch/.

Jøsang, A. (2001). 'A Logic for Uncertain Probabilities'. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9/3: 279-311, http://security.dstc.edu.au/papers/logunprob.pdf.

Jøsang, A. & McAnally, D. (2004). 'Multiplication and Comultiplication of Beliefs'. *International Journal of Approximate Reasoning*, 38/1: 19-51, http://security.dstc.edu.au/papers/JM2004-IJAR.pdf.

Kilgarrif, A. & Grefenstette, G. (2003). 'Introduction to the Special Issue on the Web as Corpus'. *Computational Linguistics*, 29/3: 333-348, http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf.

Kirkpatrick, D. (2010). *The Facebook Effect: The Inside Story of the Company That is Connecting the World*, London: Virgin.

Kiryakov, A., Popov, B., Terziev, I., Manov, D. & Ognyanoff, D. (2005). 'Semantic Annotation, Indexing and Retrieval'. *Journal of Web Semantics*, 2/1, http://www.websemanticsjournal.org/ps/pub/2005-10.

Koumenides, C.L., Salvadores, M., Alani, H. & Shadbolt, N. (2010). ‘Global Integration of Public Sector Information’. In *2nd Web Science Conference*, Raleigh, N.C., http://journal.webscience.org/303/2/websci10_submission_38.pdf.

Lanier, J. (2010). *You Are Not a Gadget: A Manifesto*, London: Allen Lane.

Lessig, L. (1999). *Code and Other Laws of Cyberspace*, New York: Basic Books.

Leung, S.-T.A., Perl, S.E., Stata, R. & Wiener, J.L. (2001). *Towards Web-Scale Web Archaeology*, Compaq Systems Research Center report #174.

López-Ortiz, A. (2005). ‘Algorithmic Foundations of the Internet’. *ACM SIGACT News*, 36/2.

Manola, F. & Miller, E. (2004). *RDF Primer*, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

McArthur, R.L. (2001). ‘Reasonable expectations of privacy’. *Ethics and Information Technology*, 3: 123-128.

Möller, K., Hausenblas, M., Cyganiak, R., Handschuh, S. & Grimnes, G.A. (2010). ‘Learning From Linked Open Data Usage: Patterns and Metrics’. In *2nd Web Science Conference*, Raleigh, N.C., http://journal.webscience.org/302/2/websci10_submission_36.pdf.

Moreau, L (2010). ‘The Foundations for Provenance on the Web’. *Foundations and Trends in Web Science*, 2/2-3: 99-241.

Morozov, E. (2011). *The Net Delusion: How Not to Liberate the World*, London: Allen Lane.

Nixon, P.G., Koutrakou, V.N. & Rawal, R. (eds.) (2010). *Understanding E-Government in Europe: Issues and Challenges*, Abingdon: Routledge.

O’Hara, K. (2009). ‘“Let a Hundred Flowers Bloom, a Hundred Schools of Thought Contend”: Web Engineering in the Chinese Context’. In X. Zhang & Y. Zheng (eds.), *China’s Information and Communications Technology Revolution: Social Changes and State Responses*, London: Routledge.

O'Hara, K. (2011). *Transparent Government, Not Transparent Citizens: A Report on Privacy and Transparency for the Cabinet Office*, London: Cabinet Office.

O'Hara, K., Alani, H., Kalfoglou, Y. & Shadbolt, N. (2004). 'Trust Strategies for the Semantic Web'. In *Workshop on trust, security and reputation on the Semantic Web, 3rd international Semantic Web conference (ISWC 04)*, Hiroshima, Japan, http://eprints.ecs.soton.ac.uk/10029/.

O'Hara, K., Berners-Lee, T., Hall, W. & Shadbolt, N. (2010). 'Use of the Semantic Web in e-Research'. In W.H. Dutton & P.W. Jeffreys (eds.), *World Wide Research: Reshaping the Sciences and Humanities*, Cambridge, MA: MIT Press, 130-134.

O'Hara, K. & Hall, W. (2008). 'Trust on the Web: Some Web Science Research Challenges'. *University of Catalonia Papers: e-Journal on the Knowledge Society*, 7, http://eprints.ecs.soton.ac.uk/16686/.

O'Hara, K. & Hall, W. (2010). 'Web Science as Reflective Practice'. In M. Cockell, J. Billotte, F. Darbellay & F. Waldvogel (eds.), *Common Knowledge: Rising to the Challenge of Transdisciplinarity*, Lausanne: EPFL Press, 205-218.

O'Hara, K. & Shadbolt, N. (2008). *The Spy in the Coffee Machine: The End of Privacy As We Know It*, Oxford: Oneworld.

O'Neill, E.T., Lavoie, B.F. & Bennett, R. (2003). 'Trends in the Evolution of the Public Web 1998-2002'. *D-Lib Magazine*, 9/4, http://www.dlib.org/dlib/april03/lavoie/04lavoie.html.

Papic, M. & Noonan, S. (2011). 'Social Media as a Tool for Protest'. *Stratfor Global Intelligence*, http://www.stratfor.com/weekly/20110202-social-media-tool-protest.

Perkins, R. & Neumayer, E. (2011). 'Is the Internet Really New After All? The Deteminants of Telecommunications Diffusion in Historical Perspective'. *The Professional Geographer*, 63/1: 55-72.

Sanchez, E. (ed.) (2006). *Fuzzy Logic and the Semantic Web*, Amsterdam: Elsevier.

Schön, D.A. (1983). *The Reflective Practitioner: How Professionals Think In Action*, London: Maurice Temple Smith.

Seabright, P. (2004). *The Company of Strangers: A Natural History of Economic Life*, Princeton: Princeton University Press.

Shadbolt, N. & Berners-Lee, T. (2008). ‘Web Science Emerges’. *Scientific American*, October 2008, 60-65.

Shadbolt, N., Hall, W. & Berners-Lee, T. (2006). ‘The Semantic Web Revisited’. *IEEE Intelligent Systems*, 21/3: 96-101.

Shadbolt, N., O’Hara, K., Salvadores, M. & Alani, H. (2011). ‘E-government’. In J. Domingue, D. Fensel & J.Hendler (eds.), *The Semantic Web Handbook*, Berlin: Springer.

Shneiderman, B. (2007). ‘Web Science: a Provocative Invitation to Computer Science’. *Communications of the ACM*, 50/6: 25-27.

Skyrms, B. (1996). *Evolution of the Social Contract*, Cambridge: Cambridge University Press.

Szomszor, M., Alani, H., Cantador, I., O’Hara, K. & Shadbolt, N. (2007). ‘Semantic Modelling of User Interests Based on Cross-Folksonomy Analysis’. In A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin & K. Thirunarayan (eds.), *The Semantic Web – ISWC 2008, 7th International Semantic Web Conference*, Berlin: Springer, 632-648.

Turkle, S. (2011). *Alone Together*, New York: Basic Books.

Victor, P., Cornelis, C., De Cock, M. & Teredesai, A.M. (2009). ‘Trust- and Distrust-Based Recommendations for Controversial Reviews’. In *1st Web Science Conference*, Athens, Greece, http://journal.webscience.org/161/2/websci09_submission_65.pdf.

Vitvar, T., Peristeras, V. & Tarabanis, K. (eds.) (2010). *Semantic Technologies for E-Government*, Berlin: Springer-Verlag.

Volz, R., Handschuh, S., Staab, S., Stojanovic, L. & Stojanovic, N. (2004). ‘Unveiling the Hidden Bride: Deep Annotation for Mapping and Migrating Legacy Data to the Semantic Web’. *Journal of Web Semantics*, 1/2, http://www.websemanticsjournal.org/ps/pub/2004-15.

Warf, B. & Grimes, J. (1997). ‘Counterhegemonic Discourses and the Internet’. *Geographical Review*, 87/2: 259-274.

Wilks, Y. & Brewster, C. (2009). ‘Natural Language Processing as a Foundation of the Semantic Web’. *Foundations and Trends in Web Science*, 1/3-4: 199-327.