The Semantic Web: Legal Challenges

The Semantic Web is about making data smarter and linking that data up. Smarter, interlinked data means data that is easier for computers to access, interpret, process and re-use. The promise of the Semantic Web is of a vast network of interconnected nodes of data, accessible to any computer and application connected to the internet. This is the vision of a more powerful, better integrated web of data which backers of the Semantic Web see as the core aspect of Web 3.0, the next generation of the World Wide Web.

The technologies that make up the Semantic Web have not yet fully matured, but they have reached a stage in their development where it now makes sense to start taking a hard look at the practical, legal issues to which their implementations are likely to give rise. The technologies included under the umbrella of the term ‘Semantic Web’ are primarily standards for encoding smart data. The two core standards, RDF and OWL, were adopted as recommendations by the World Wide Web Consortium (W3C) in 1999 and 2004. While Web 3.0 has yet to become a reality, a growing number of vendors make use of these standards to provide enhanced functionality:

Yahoo! has integrated semantic elements into its search engine; bbc.co.uk is an enthusiastic backer of the technologies, using them to improve the cataloguing and organisation of its vast database of content; Oracle’s flagship database now comes with an RDF option – to name but a few.

What is the Semantic Web?
The cluster of standards that makes up the Semantic Web is complex, and we will provide no more than the briefest of overviews. The foundations of the Semantic Web, RDF and OWL, are simply standards for representing and structuring data. What is novel about these standards is the evolutionary leap in the organisation and processing of data that they enable.

The technologies that make up the Semantic Web have not yet fully matured, but they have reached a stage in their development where it now makes sense to start taking a hard look at the practical, legal issues to which their implementations are likely to give rise. The technologies included under the umbrella of the term ‘Semantic Web’ are primarily standards for encoding smart data. The two core standards, RDF and OWL, were adopted as recommendations by the World Wide Web Consortium (W3C) in 1999 and 2004. While Web 3.0 has yet to become a reality, a growing number of vendors make use of these standards to provide enhanced functionality:

Yahoo! has integrated semantic elements into its search engine; bbc.co.uk is an enthusiastic backer of the technologies, using them to improve the cataloguing and organisation of its vast database of content; Oracle’s flagship database now comes with an RDF option – to name but a few.

What is the Semantic Web?
The cluster of standards that makes up the Semantic Web is complex, and we will provide no more than the briefest of overviews. The foundations of the Semantic Web, RDF and OWL, are simply standards for representing and structuring data. What is novel about these standards is the evolutionary leap in the organisation and processing of data that they enable.

One of the driving ideas behind the Semantic Web is to create a web of data instead of, or in addition to, the current ‘web of documents’. The expression ‘web of documents’ refers to the current World Wide Web, which is essentially a vast network of hyperlinked documents, primarily coded in HTML. While we human beings can easily navigate these documents and make sense of their contents, the data which most web pages contain is not structured in a way that is easily machine-readable. A table on a webpage that sets out the highest mountains in the world encoded in HTML means very little, on its own, to a computer: the data itself does not give a computer any clues that would enable it to know that ‘K2’ is the name of a mountain, or that ‘8,611’ is a measurement of that mountain’s height in meters. What Semantic Web standards offer is a means of encoding that data so that the data itself provides pointers as to what the raw data is about – what it means.

RDF and OWL: Describing Resources
The basic building block of the Semantic Web is the ‘resource’. A resource is essentially anything that can be identified: things on the internet (a web page, a blog post, etc.) and things beyond the internet which are referred to on the internet (a product, a book, a person, a concept). Resources are identified by means of Uniform Resource Identifiers (URIs). URIs resemble URLs (web addresses) in form, but their primary function is not so much to act as locators for resources (though this can also be the case, and frequently is), but rather to provide a unique name for each resource across the internet.

RDF, the Resource Description Framework, is the foundation layer of the Semantic Web: it provides a framework for making assertions about resources, called ‘statements’, or
'triples'. The latter name derives from the structure of an RDF statement, which is always in three parts:

1. **Subject**: the thing the statement describes (the resource);
2. **Predicate**: a property which is being asserted to belong to that thing;
3. **Object**: the value of that property (which can be another resource).

Thus I might describe a resource (the Subject), identified, say, by the URI 'http://www.example.com/books/the-wealth-of-nations', as having the property ‘title’ (the Predicate), with the value ‘The Wealth of Nations’ (the Object). This means that the resource identified by the URI (which is a book), has the property ‘title’, and that title is ‘The Wealth of Nations’. The network effects start to kick in when multiple related statements are made, creating a web of interconnected ‘triples’: I might create further triples stating that this URI has a property ‘author’, which points to another URI identifying Adam Smith, which can also be the value of the ‘author’ property for other resources identifying other books by Adam Smith, and so on.

The other major pillar of the Semantic Web is OWL, the Web Ontology Language. The word ‘ontology’ is taken from philosophy, in which it describes the study of **being**, addressing such questions as what entities can be said to exist, how such entities are related to each other, how they can be classified, grouped and distinguished. This is also what a web ontology does: using the formal language provided by OWL, a web ontology describes a set of ‘concepts’ and the relationships between them. These vocabularies of concepts can then be used with RDF to make statements about particular instances of things. Because OWL enables developers to specify the formal relations between concepts with a great degree of formal rigour, OWL-enabled applications are capable of drawing complex inferences from appropriately structured data.

A growing number of such vocabularies (or ‘ontologies’) have been developed and are in increasingly frequent use, including:

- the 'Dublin Core', a vocabulary used to describe information resources, eg ‘title’, ‘creator’, ‘publisher’, ‘language’, etc.;
- the Friend of a Friend ontology (FOAF), which is used to describe people and their social network (see below);
- the Semantically-Interlinked Online Communities ontology (SIOC), developed by DERI at NUI Galway, to describe information from online communities, such as message boards, wikis, blogs, etc.

What RDF and OWL achieve, which sets them apart from existing data standards, is that they place meaning directly within the data, rather than within the code of the program which processes the data, and this is what is meant by making data **smarter**. Because this is achieved by means of web-based URI references, the resulting data is not only **locally** smart, but is connected into a vast network of smart data across the internet. This is made possible by another key feature of these data formats: that they are graph-based. Drawing on the field of mathematics known as graph theory, they enable data to be structured in networks of nodes which can be easily merged – something which has been difficult to achieve in earlier hierarchical data formats (such as XML). It is this ability to merge which gives the Semantic Web the potential to evolve and grow into a global online web of data.

We should point out, before embarking on our analysis of the legal issues, that much of what Semantic Web technologies achieve is uncontroversial: providing smarter standards for encoding data, in itself, is probably no more legally controversial than using existing data standards such as XML. However, there are aspects of the Semantic Web, especially in its applications, which do give rise to identifiable legal challenges, and it is these that we will focus on.

### Data Protection in the Web of Data

Back in 2001, in their influential article on the Semantic Web in Scientific American, Tim Berners-Lee, James Hendler and Ora Lassila described a speculative scenario which aimed to demonstrate the benefits of smarter, integrated data. The scenario involves a woman named Lucy who uses her handheld web browser to generate a plan for medical treatment for her mother, drawing on data which includes her mother’s medical prescription, insurance details and home address. While the scenario is impressive from a technical perspective, it is also bound to set alarm bells ringing with data protection lawyers.

Under EU data protection law, personal data is defined as information relating to an identified or identifiable natural person, the data subject. The Directive imposes a range of obligations on data controllers (persons who determine the purposes and means of the processing of personal data) and data processors (persons who process personal data on behalf of the data controller). These obligations include obligations to ensure that the personal data is processed fairly and lawfully; that the data is collected only for specified, explicit purposes and is not processed for any incompatible purpose; that the data should not be kept for longer than is necessary; that appropriate security measures are taken against unauthorised access, etc.

Furthermore, where personal data is obtained directly from the data subject, the data subject should generally give consent for the processing of his or her data. Where the data is not obtained directly from the data subject, the data still has to be processed ‘fairly’, which requires that, insofar as practicable, a number of conditions are fulfilled, especially relating to informing data subjects of the identity of the data controller and the uses to which the data will be put.

In addition to these already onerous obligations, further stringent requirements are imposed where the data constitutes ‘sensitive personal data’, which includes data relating to the race, political opinions or religious beliefs of the data subject; membership of trade-unions; data concerning the physical or mental health or sexual life of the data subject; and data relating to the commission of criminal offences and any related proceedings.

One of the main aims of the Semantic Web is to make data easier to process and re-use: the idea is that the data made...
available will be accessed over the internet, processed and integrated with other data by a vast array applications for any imaginable purpose. What becomes of the protection of personal data in such an open, universally accessible web of interlinked data?

A first response might be that all of the data on the Semantic Web will be public data, implying some type of universal consent: if somebody has gone to the effort of encoding data using Semantic Web technologies and making it available on the internet, it is arguable that such a person has effectively consented to his or her personal data being subjected to broad uses arising from semantic technologies. Even if the data does contain personal data within the meaning of data protection legislation, surely the person posting the data can be assumed to have consented to the further processing of his or her data.

This argument is not without its merits, but it ignores several important points: first, the future of the Semantic Web does not lie in specialists ‘manually’ encoding and posting data to the web, but rather in automated encoding of data into Semantic Web formats by applications. This raises the question of whether the data subject who uses that application really understands how widely available that data may become as a result, and therefore whether he or she is really giving informed consent to the processing of the data. Should data capture applications that automatically encode personal data be required to alert users to their existence in the same way that cookies are regulated under the Data Protection and Electronic Commerce Directives? Secondly, even where a data subject consents when first making his or her data available to a data controller, this does not exempt other data controllers who make use of that data from the requirements of fair processing (eg notifying the data subject of the data controller’s identity, intended uses etc.). Finally, the data may well include information about people other than the person who is making the data available: that person cannot consent on behalf of the other data subjects involved. Here again, questions may arise as to whether this data is obtained fairly.

One of the reasons data protection is of such concern is that semantic applications are likely to prove far more effective than conventional search engines at piecing together scattered but interrelated pieces of data, potentially recreating detailed profiles of data subjects at the click of a mouse. If not carefully handled, the integration of personal information into the web of data would be sure to prove a boon to spammers, identity thieves and other fraudsters.

Friend of a Friend of a Friend

The FOAF (Friend of a Friend) ontology is particularly interesting from a data protection perspective, because a FOAF profile is essentially a bundle of personal data. Using FOAF, I can create a data file which I can then make available on my homepage, or indeed anywhere on the internet, which sets out information about me, such as my name and my e-mail address, along with information about people I know.2 Because FOAF is a
Semantic Web ontology, the idea is that each person I know should be uniquely identified by means of a URI, and that they might have a FOAF profile of their own, which in turn lists the people they know. In theory, an application could reconstitute the entire graph of every single person with a FOAF profile who is ultimately connected to me through a chain of FOAF profiles, no matter how many degrees removed, along with all of the personal details they have included in their FOAF files.

Any realistic implementation of Semantic Web technologies which involves personal data, especially sensitive personal data (such as Lucy’s mother’s medical records), is going to have to provide the means to regulate access to the data. This is an issue which the Semantic Web community is acutely aware of, and a range of solutions involving authorisation and access levels are in development. Because semantic data is smart data, it should be possible to integrate, within the data itself, information about who should be allowed to access the data, under what conditions it should be transferred, etc.

A number of solutions to the data sharing problem are emerging. As mentioned earlier, an RDF structure consists of three elements: Subject, Predicate and Object. Adding a fourth element, Context, allows the data provider to include information on the provenance of the statement, which may assist in determining permissible uses of the data. In addition to this, researchers in DERI are looking at ways of attaching machine-readable licences to RDF statements. Attaching licences in this way might enable rights in the data to be determined automatically.

The Risks of Inference
One of the most exciting aspects of Semantic Web technologies is the possibility for greatly enhanced processing of data and in particular the ability of Semantic Web applications to draw inferences from data by exploiting rich formal meta-languages. This promises a generation of applications that can dig much deeper into data than, for example, the keyword search/document retrieval of current search engines. By drawing inferences from structured data throughout the web, semantically-enabled applications can generate new statements. With the ability to generate new statements, however, comes the risk that those statements may be false or misleading.

There might be some merit in drawing up data protection guidelines specifically tailored to the use of Semantic Web technologies. Perhaps this is something which the Article 29 Working Group should consider placing on its agenda. Better still, from the developer’s perspective, would be to ensure that implementations of the standards are coded in such a way as to actually prevent breaches of data protection principles. For instance, if I set up a Facebook-type social networking Web site which automatically generates FOAF profiles for users, I should ensure (a) the user is fully aware of what may become of his or her personal data and consents to this; (b) that the FOAF files generated contain only minimal information about other persons in my social network (eg URIs only); and (c) that the availability of the resulting FOAF profiles or other semantic data files is carefully controlled. These safeguards should be achievable by means of a combination of the underlying code and information provided on the site (especially the privacy policy and user interface information prompts).

The plain English text of the source data which this statement is extracted from is very careful to qualify this famously controversial allegation, eg ‘according to three United States government investigations …’ or ‘the Warren Commission concluded that …’, without stating directly that Lee Harvey Oswald killed Kennedy. The nuances that these qualifications provide are stripped away by the extraction process, leaving
only the blunt assertion. As any newspaper editor will confirm, it is precisely these types of nuance and qualifications which save publications from many a defamation suit. Clearly this type of error or distortion is a source of legal risk.

Though the defamatory statement is strictly speaking generated by a ‘machine’, responsibility for defamation attaches to the person deemed to be the ‘publisher’ of the defamatory statements. It may not always be entirely obvious who this person might be: for a service available over the internet, it would in all likelihood be the company operating the application which generates the results, though in certain circumstances liability may also attach to other parties. In defamation law, any party that re-publishes defamatory material is also deemed a publisher and liable in defamation. If one application were to draw and publish online a defamatory statement which was inferred from an incorrect, defamatory statement generated by another application, the person responsible for the operation of the infringing application could also be exposed to a defamation suit.

It might be argued that, because the statement is generated by a machine, it is unlikely that it would damage the reputation of the plaintiff in the eyes of a reasonable person, because a reasonable person is unlikely to regard machine-generated statements as equivalent to human judgments. Recent caselaw regarding the juxtaposition of elements in automatically generated web page content demonstrates that automatically generated content may indeed give rise to legal risk: for example, a Dutch news portal was successfully sued earlier this year because the Google-generated summary of one of its articles gave the misleading impression that the plaintiff was bankrupt. This could be further compounded by the fact that a reasonable person may not always be on notice that information has been automatically generated: if expressed using natural language processing technologies, statements generated by a semantic application may give a convincing impression that they were in fact authored by a sentient human being.

Defamation is only one of the concerns that the possibility of error in inference gives rise to: liability in negligence could arise where an application produces incorrect information which the user relies on to his or her detriment; contractual issues of misrepresentation or mistake could arise where an automated service draws incorrect inferences, leading a user to enter into a contract he or she would otherwise not have entered into – think, for example, of an online price-comparison Web site which produces an incorrect comparison.

The fact that code can have bugs and that this can lead to errors is nothing new. What is new, however, is the inferential power of Semantic Web technologies, and the expectations to which these can give rise. For developers of semantic applications, this possibility of error, and its legal implications, should be kept firmly in mind. Terms and conditions of use should prominently disclaim any responsibility for the accuracy of the information provided, and automatically generated data should be clearly identified as such.

Conclusion

The Semantic Web is an exciting developing area that appears to be gaining the momentum that will enable it to deliver the next step in the evolution of the web. The direction of that evolution presents real challenges to the current legal framework governing the processing of information, based as it is on concepts of data and information flows that have their roots in the pre-internet era. It has been remarked that true innovation often depends on a little law-breaking. Certainly, some of the most widespread internet services in daily use today do not sit entirely comfortably with the legal structures that regulate them, but this has not (yet) proven a major impediment to their success.

However, the legal challenges faced by the Semantic Web are not simply a matter of innovation versus inflexible regulation. They touch on issues that are among the central concerns of web users: privacy and the reliability of information. The success of the Semantic Web will in part depend on the ability of those in the field to address those concerns, while enabling the technology to flourish. Perhaps one of the most exciting prospects for Semantic Web technologies lies in the possibility that many of the legal challenges which they give rise to may themselves have semantic solutions. Averting the legal risk may not so much require the intervention of lawyers and regulators, but rather making the smart data smart enough to control its own legal effects.

Brian Harley is a commercial lawyer at Mason Hayes+Curran with a particular interest in emerging technologies.

Philip Nolan is the head of the Commercial Department at Mason Hayes+Curran and a leading Irish IT lawyer.

Liam Ó Móráin is a business development consultant to DERI.

Mark Leyden is a research fellow at DERI.

Endnotes


2 Curious readers can generate their own FOAF file at www.loddds.com/foaf/foaf-a-matic

3 Optimized Index Structures for Querying RDF from the Web, A. Harth, S. Decker, Digital Enterprise Research Institute (DERI)

4 http://www.powerset.com ; see also the DERI project http://sig.ma, which is currently in alpha testing.

5 Site aansprakelijk voor Google-indexering, De Telegraaf, 14 May 2009; see also http://www.24oranges.nl/2009/05/17/site-convicted-for-gooles-%0Aautomatic-abstracts/