Guru Interview:
Sir Timothy Berners-Lee, KBE

The father of the World Wide Web explores the development and potential power of the Semantic Web.

Interview by Sarah Powell

Professor Sir Tim Berners-Lee KBE is the inventor of the World Wide Web, and founder and director of the World Wide Web Consortium (W3C).

The consortium coordinates Web development worldwide together with teams at MIT, The European Research Consortium in Informatics and Mathematics (ERCIM), and Keio University in Japan. He is a senior researcher at the MIT Computer Science and Artificial Intelligence Laboratory where he holds the 3Com Founders Chair. He also serves as Professor of Computer Science at the School of Electronics and Computer Science (ECS) of the University of Southampton.

Since 1995 Tim Berners-Lee has received a raft of international awards, nominations, honorary degrees and fellowships including a MacArthur Fellowship in 1998 and, in 2002, the Japan Prize from the Science and Technology Foundation in Japan. In 1997 he was awarded the Order of the British Empire (OBE), and in 2004 he received a knighthood (KBE) for services to the global development of the Internet via his invention of the Web. In June 2004 in Helsinki Sir Tim Berners-Lee was awarded the first ever Millennium Technology Prize 'for outstanding technological achievements that directly promote people's quality of life, are based on humane values, and encourage sustainable economic development'.

Tim Berners-Lee was a fellow at CERN, the European Particle Physics Laboratory in Geneva, Switzerland, in the late 1980s and early 1990s, which is when he developed the World Wide Web. He made available the first programs for browsing the Web within CERN in December 1990, and then publicly, free for anyone to use, in the summer of 1991.

The history, development and vision behind the Web are described in his book *Weaving the Web* (Harper, San Francisco, 1999).

How would you describe the Semantic Web and how does it build on the World Wide Web to enable computers not only to make links but also to extract meaning from disparate information to create a web of data?

Sir Tim Berners-Lee:

Computers can't 'understand' data in the way that people do. The Semantic Web is not about 'understanding' data; it is about putting data onto the Web to make it available so that we can access and use it. There is a tremendous lot of data out there which is why some people have called it the Web of data or the 'Deep Web', meaning that it is not really accessible because we can only probe it through websites which have databases behind them. But documents, bank statements, rolls of films and so on are all data files too. So the Semantic Web is about getting data onto the Web, and data in the 'Deep Web' can then be exposed with a language. We have RDF, or Resource Description Framework, as the data language and we have developed a query language called Sparql (pronounced 'sparkle').

In terms of sharing information and mimicking the human association of ideas through linking information, the World Wide Web builds on the ideas and work of people such as Vannevar Bush and Doug Engelbart. The hypertext Web brings us a common information space – global information at the click of a mouse. The Semantic Web goes further, providing a web of data but also allowing for machine analysis of RDF data content.

How does RDF work? Does it rely on pattern recognition?

Sir Tim Berners-Lee:

RDF language is about interoperative data. Unlike pattern recognition, which is about trying to discover pattern where it's not explicit, the Semantic Web is about explicit relations.

We can look at pattern recognition and about trying to understand meaning from human discourse and so on, which some fields of artificial intelligence do, but that's not what the Semantic Web is about. The Semantic Web is about the interoperability of reasonably well-defined data where you have well-defined relationships.

The Web ontology language, or OWL, allows you to find terms like 'parent' and 'earth mother' – they're relationships between people – or 'author' – the relationship between a person and a document. Or it might be something such as the relationship between a protein and a gene, and so on. These are the sort of things we're usually searching for and they are captured in

part but they're not all connected together. For example, I can't formulate a query to find all the papers written by friends of people who work on a given project. This should be a simple data query. The problem is that all that data is in databases which are disjointed, and often in different data formats.

The World Wide Web has been immensely empowering in terms of promoting access to information, raising awareness, boosting understanding and providing a forum for free expression. Many facets of our lives have been revolutionized as a result. Is this the 'magic' you predicted in Weaving the Web and what sort of applications do you envisage for the Semantic Web?

Sir Tim Berners-Lee:

Well, some of the things that have happened on the Web do seem magical in that, when people get together and collaborate, it is magic – and really unexpected things can happen. Every now and then I get an e-mail from somebody who has found that they've managed to do something with the Web that they really didn't expect. That is very nice to hear. But there are also many times when you're using the Web and it's crazy – it's clear that the community isn't helping you. For example, you might be looking at a Web page about a meeting and all the information you need about time, place, people etc. is there, but you will have to copy all the information from the Web page into your address book by hand if you want it to be instantly accessible. This is such a waste of time. Our hope is that the Semantic Web will allow for integration of data-oriented applications as well as document-oriented applications.

The power of the Semantic Web is its ability to link things together, to connect things. Anyone can put a data application on the Web and there are some wonderful websites out there, but the need is for them to act together. Just cast your mind forward and imagine you have a company. Imagine that, when you are making purchases, all your suppliers have put data on the Web about the parts they sell and their pricing and delivery details. With all this information and details on part compatibility available on the Semantic Web, I can write a program to find the best fit for a new part if I break one and need a replacement.

What's also exciting about the Semantic Web is its potential for unexpected, serendipitous re-use of data, i.e. when somebody uses that information for a completely different purpose. Again, imagine you have all this information about different companies selling different parts. When people also put their pricing information on the Web in an analyzable form, a researcher will be able to pull up the prices of various parts to analyze the variations across the USA. Such

research could be really interesting – you could, for example, be investigating the costs of transport. Or perhaps your company is facing a crisis, a fire at a plant, and you fear it will be unable to produce a particular product. You might want to identify which customers will be affected so as to organize for members of staff who live closest to their plants to visit them individually. It would involve taking data from the HR database, from the customer database, from the orders system and the production system. Normally it would take ages to try to put together a system that integrates all that data. In the future it could all be on the Semantic Web, and accessible using Sparql. The whole point of the Semantic Web is to connect different applications so that people can just ask a question and then access and navigate around all that data.

In your book you write that to foster trust and confidence in Web content and promote quality you would like to see the use of endorsement techniques similar to the PICS protocol to express other subjective notions such as academic quality. Is this occurring to any extent?

Sir Tim Berners-Lee:

I think it is coming. One thing that's already happened is that, generally, a hypertext link is considered to be an endorsement to a certain extent. Because Google, for example, considers a link as an endorsement, another interesting development is that now you can choose when you make a hypertext link whether you want it to be used or ignored as an endorsement. If, for example, you want to refer to something with which you disagree, you can put an attribute onto a link to ask search engines not to take it as an endorsement.

There are numerous endorsement-based websites out there. When you think about it, the whole blogging community works as a form of endorsement. Bloggers point at other people's blogs, and this endorses them. Then you can take the track-back system of lists of blogs that have pointed to your blog as a measure of kudos. All this and much more is already happening. The Internet Content Rating Association, ICRA, has also launched a drive to create some sort of RDF-based labelling system on the Web – this would be in the PICS style but different from PICS which is now rather out of date.

What sort of things would be labelled?

Sir Tim Berners-Lee:

Well, the whole point about endorsement on the Web is that all kinds of things can be labelled. For example, there's a move to endorse websites as being suitable for mobile phones. Here at W3C we have launched the Mobile Web Initiative (MWI) to improve the experience of browsing the Web with a mobile device. We're calling for endorsements about sites that are accessible, use the standards and so on. There is also interest in labelling sites to indicate whether they're suitable for children or not.

There has been great enthusiasm for labelling but we're not talking about government labelling of sites. Sometimes it's labelling by the site itself, or else it's labelling by a third party that makes the decisions. There are many different labelling systems. It's not a rating system such as that for films.

> ### *"The threat at the moment in the USA comes from telecommunications companies that have noted the vast profits made by Google and are threatening to charge more specifically for each customer that they connect."*

W3C's role is to develop open Web standards and guidelines to ensure the long-term technical evolution and ongoing network neutrality of the Web. Speaking at WWW2006, you warned of a threat to charge for different levels of online access. Where is this threat coming from, what are the implications, and how is it being combated?

Sir Tim Berners-Lee:

While we may pay for different service levels, e.g. we pay more for a higher bandwidth, the important thing about the Net is that if we both pay for a certain level of service, then we can communicate at that level no matter who we are. We pay to be able to connect to a certain bandwidth and that's all we have to do. It's up to our ISPs to ensure that the interconnection is done. This is how it has always been done. The threat at the moment in the USA comes from telecommunications companies that have noted the vast profits made by Google and are threatening to charge more specifically for each customer that they connect. As an example of the potential impact, imagine that my MIT ISP decided to block Emerald and then went to you and said: 'Look, you're delivering interesting content, but to deliver specifically to MIT you'll need to join our partnership programme and we'll help you deliver your content even better to MIT, e.g. delivering video or audio as well'.

The telecommunications companies are trying to argue for the right to do that – to block and selectively allow high value-added connectivity so that effectively they would control the video sites their customers could access. They want to stay in the cable TV market because they have the ability to negotiate particular sources of TV streams. But TV over the Internet should be a different arrangement. Emerald should be able to start up on a TV station, buying up bandwidth for a few people to be watching it at any one moment. For your customers, that would be one of a huge choice of different TV channels that they could access.

As far as I can see at the moment this threat has only emerged in the USA. There have been battles in Congress and channel TV companies have spent massive amounts of money lobbying and putting ads on the Internet to try to pretend that they're being more open than the people who are demanding the neutrality of the Net. The TV companies are putting forward all kinds of ridiculous arguments. In response there has been a public outcry, in fact a huge outcry from people at both ends of the spectrum, ranging from MoveOn.com, a more or less democratic or left-wing political body, to gun-owners and some right-wing religious organizations. Many of these people realize they risk seeing their websites blocked because the media companies don't really approve of them. In fact anyone who worries that their views might not be those of the mainstream media is concerned by this.

What challenges do you face in ensuring browser security, given the dangers of phishing, viruses and spam?

Sir Tim Berners-Lee:

This is a complicated question to answer as there are so many of these. I see the most difficult challenges as being in the media interface with the human being. Phishing relies on the human being hoodwinked, so it is a question of making a browser, an e-mail client, which more transparently indicates what is going on, i.e. so that it is clear who is sending messages or who is running a website.

W3C had a workshop on this recently and we have some work coming up on it. Meanwhile there is plenty of cryptography out there. When it comes to traditional security, this involves making crypotographic, coding schemes which are more difficult to break. But that's not the issue here. The issue is one of trust management. It is how to design a system so that the user can recognize when a website that claims and seems to be their bank is not...

Some quite simple things can be done to foster trust. In addition to having a little padlock to indicate a secure site, a browser could also display the name of the holder of the certificate. There are also more complex things to help the user to manage the trust on their system, telling the computer what information to trust. We've been talking about these issues for a year or so and there are a number of challenges to be addressed relating to building the interface to all systems. When you're doing something in the physical world and the hairs on the back of your neck rise, this indicates that some part of your brain is making you suspicious. But it's very difficult to write down what it is. We have to find a way of making it much more explicit so that the computer can help with it.

You have described your role at W3C as 'facilitator of consensus'. How does the W3C reconcile the tensions between the open source focus of what you have described as 'philosophical engineering' and the commercial interests of many of your members, and to what extent is software patenting a problem in development of the Web?

Sir Tim Berners-Lee:

There are a number of tensions given the role of facilitating consensus, visualizing it, arriving at a common agreement despite people coming from different initial positions. Just talking about a problem when different participants have stakes in different outcomes gives rise to tensions. But the consortium works in an area where there are innumerable win-win situations. There is, for example, a win-win situation in that a successful standard will typically create a large, new market. This means that everybody involved makes an effort, understanding that.

In the early days of W3C when we were working on the P3P technology for privacy, a participant claimed to have patented part of the developing standard. This was a real problem and discussions went to a very high level. It also forced W3C to take on the issue of patents and Web standards, and we put together a Working Group with diverse representation to develop a policy – one that received public review throughout its development. As a result of this experience, W3C developed its innovative Royalty-Free Patent Policy to help ensure that core Web standards can be implemented at no cost. Although it took several years to develop, it was adopted in 2004 and is now well established and well respected.

Large companies that you might have thought of in the past as being champions of IP did the maths and realized that it's very important to have royalty-free standards. They could see what

had happened to things that were not royalty-free: how these would take off as a very limited market; and how things which were more open were just much livelier, and a foundation for huge, new unexpected things to be built on top of them.

That message has got across now so that the large companies very much understand this. In fact they sometimes bring work to us at W3C, saying that they don't want to do it anywhere that doesn't have this policy, because they don't want to work on something and then find that it's not sufficiently open. So while there's always a possibility of some person trying to maintain that they invented the whole thing ten years ago, in practice the players who are in a position to take such a stand are generally involved in making the standards and sharing the IP. So the patent policy has been a great success, even though every now and again we still meet problems that need addressing. A few companies had to do some soul-searching before adopting it and there are some companies that are still unhappy, but the vast majority of the companies that matter are very pleased with it.

You have emphasized that the Web is a social creation, its value deriving from the opportunities it presents to tap into and share knowledge and communicate freely across a common information space. How are the social and business benefits of collaborative open source initiatives such as the Web reconciled?

Sir Tim Berners-Lee:

The open source world and the commercial world are both very important. They always have been and they have to co-exist; we need both. When the Web started, of course, it was all open source. There were no products and the response we got from a lot of commercial users was that they didn't really want to rely on something for which they couldn't pay for support. They'd use a browser when they could buy one. So the starting of Netscape Communications was an important step then.

When we look now to the Semantic Web there's a broad base of open source. There's a set of start-ups, producing Semantic Web technology, and there are large companies that are used to handling data in large quantities and are now producing Semantic Web-based products and connecting their products to the Semantic Web. Both are important and both will continue to be important.

The open source side is extremely important for the creative, for the academic, for computer science and websites. As a discipline you need open source platforms on which people can test out and develop new ideas.

How close is your dream of 'one Web for everyone, everywhere, on everything' to realization?

Sir Tim Berners-Lee:

The dream is not yet realized because we don't have a collaborative space. We're only starting to get our data on the Web and we have a very small proportion of people in the world connecting to the Internet. There are many, many ways in which we have a lot of work to do. □

Republished from Emerald Now, October 2006.