The future of the search engine: "Search and you will find". An interview with Frank Van Harmelen

Frank Van Harmelen[1]

[1]Frank.van.Harmelen@cs.vu.nl http://www.cs.vu.nl/~frankh/

AI Department, Division of Mathematics and Computer Science, Faculty of Sciences, Vrije Universiteit Amsterdam, de Boelelaan 1081a, 1081HV Amsterdam , The Netherlands

Abstract. The paper presents, in form of interview, the author's view on the Semantic Web and its development in the future.

Sommario. L'articolo presenta, in forma di intervista, le opinioni dell'autore sul Semantic Web e sui suoi sviluppi futuri..

The Web is sometimes described as an enormous library where someone threw all the books in a big pile on the floor, and where you have to wear a blindfold when searching for the right piece of information. Search engines such as Google are of course a big help, but sometimes even they are no match for the chaos that we find at the current World Wide Web. Will this ever change? In this interview we are trying to cast on eye on the future of the search engine.

The World Wide Web Consortium (or: W3C), the organisation that promotes the development and the evolution of the World Wide Web is well aware of the rather unruly grow of the Web, and all the problems that this brings.
Therefore, they have already been working for some time on solutions to this problem. That solution is called the Semantic Web, which should form the next generation of the current World Wide Web. We asked Frank van Harmelen, professor in Computer Science at the Vrije Universiteit in Amsterdam, and closely involved in the development of the Semantic Web, for an explanation.

Q: "Let's begin with the most obvious question: what is the Semantic Web?"

Van Harmelen: "The best way to explain that is to look at the shortcomings of the current World Wide Web. On the one hand it is a huge success - nobody could have predicted ten years ago that the Web would be such an enormous influence on our daily and lifes, both personal and professional - but on the other hand it has a number of important shortcomings. The current Web is only usable if you speak English (or some other widely used language), and if you can recognise images and pictures.

http://www.dif.unige.it/epi/networks

Networks 2: 1-6, 2003
 http://www.swif.uniba.it/lei/ai/networks/

People are very good at this, but computers can't do that at all. Computers can't deal at all with the current Web, at least not with its content. As a result, we currently get only very little support from our computers when we are looking for information on the Web. The only job that our expensive PC's are doing for us is to move information from one location to another, and then to display the information on your screen. But to understand that information, to combine it, interpret, select and judge it, all that is left entirely to the human user. The computer cannot help us with that, simply because it doesn't understand what all these Web pages say. So, what about the Semantic Web? Well, the main idea behind the Semantic Web is that we try to extend the current Web with additional information to make it possible that computers do understand the content of web pages. Of course, we are not going to remove or replace the current Web. The Semantic Web is an extension, an additional layer on top of the already existing Web. It simply means that we have to extend the information in the current pages to make them understandable to computers. And we are well on our way to achieving that."

Q: "Can you tell us more about how that is being done?

Van Harmelen: "We have developed a number of languages, that can be processed by computers, and which can describe to the computer what the content is of a particular web-page. For example, you could state in such a language that there is something called the 'Vrije Universiteit in Amsterdam', that there is somebody called 'Frank van Harmelen', and that these two objects have a specific relation, namely 'is employed by'. You could also define the concept of a 'building', that a specific building 'is part of' the Vrije Universiteit, and that Frank van Harmelen 'works in' that building. Of course, all these relations must be defined: the relation between me and the building ('works in') is fundamentally different from the relation between the university and that building ('part of'). When you describe all of that in one of the languages that we have constructed, the computer will understand what you are looking for, simply because you explained this to the computer beforehand. In that way, the computer could already give you much more support on your search for a specific piece of information."

Q: "So then, the entire system depends on how you provide the computer with all this information?"

Van Harmelen: "That is indeed correct. Let me give you another example: suppose you are looking for the work-address of Frank van Harmelen. Maybe you won't find that with Google, because I was too lazy to put this information on my website. But if the computer knew that I work for the Vrije Universiteit, and it could find the address of the university, then it could deduce that this was a useful address to return in answer to your question. Of course, this all depends on the background information (the knowledge) that the computer has been provided with. So, everything will indeed depend on the quality of what we call the "ontology": a collection of terms and relations between these terms."

Q: “In other words, the quality of the ontology determines the quality of the support that the computer can give you.”

Van Harmelen: “Exactly. You could regard an ontology as a structured way to represent the meaning of words in a given domain. To come back to the example: to make that work, you must explain to the computer what a university is, what an employee is, what the relationship between these two is (for example, that the work address of the employee is the address of the University), etc. All of this information together is called metadata. And that is exactly what is needed to create the Semantic Web. Without those metadata there will be no Semantic Web.”

Q: “But where is all those metadata going to come from?”

Van Harmelen: (smiles) “That is indeed the question I get asked most often if I give presentations about the Semantic Web. Of course, users will not write these metadata. If we look at the origins of the current Web - say the first hundred thousand pages or so - then those pages were still written by people, using their keyboards to type up HTML pages. But of course that is no longer the case. We didn't get to more then 3 billion pages on the current Web by typing them in! Instead, most pages are being generated from databases, constructed by computer programs, etc. Well, in the future those databases and computer programs will not only generate the HTML (for human consumption) but also the metadata (for computer consumption). A simple example is the website of Amazon. Of course, that website is simply generated out of a database. All the information from that database is turned into HTML pages, so that we, the human users, can read and understand them. But of course, Amazon could also use the same database to generate the metadata in a language that is accessible to a computer. And that would allow my personal shopping agent to assist me with search for books or music that suit my personal tastes, again described in the form of metadata. So these databases are already one big source of metadata. Another important source of metadata are specialised programs which can understand - in a superficial way - natural languages such as, for example, English or Italian, and which can generate metadata out of natural language sentences. Those programs already exist, and there are already companies earning their money with those programs. To sum it up: metadata will be mostly generated by machines are automatically or semi-automatically.”

Q: “But shouldn't those metadata be standardised in some way in order to allow for exchange of the metadata?”

Van Harmelen: “That is indeed an important point. For example, when I was talking about ‘employee’, and someone else talks about ‘personnel’, then the computer needs to know that these are the same concept, so that, when it is searching for employee, he should also look for ‘personnel’. That is exactly what we hope the Semantic Web will do for is. The current search engines are mostly limited to matching strings. Granted, they are a bit more clever then that, but fundamentally they rely on finding the same sequences of digits across web pages. The use of ontologies should fix that: we are no longer restricted to matching on the string ‘employee’, but the concept

‘employee’ and its relation to other concepts will be defined in an underlying ontology.
Of course, when someone else uses the word ‘personnel’, he should refer to a similar ontology, and these two ontologies must be linked to each other. Only then will the computer be able to realise that these two concepts are the same, and that one can be used while searching for the other. In order to get that working, you really need the standardised metadata languages, as provided by W3C.”

Q: “Can't the computer make that link for itself? Don't we have the technology to do that?”

Van Harmelen: “That is indeed a very hot research topic. We can indeed already do this in experiments using carefully chosen test domains, but we cannot yet do it on the big open world of the real Web. I expect that this ontology matching problem will generate an entirely new commercial market.
Already, companies are selling ontologies. For example, they might have a very large commercial ontology, describing among others terms such as ‘employer’, ‘employee’, ‘personnel’, ‘address’, ‘product’, ‘price’, together with a description of the relationships between these terms. When you pay them, you are allowed to link to terms in this ontology. They then take care of linking their ontology with other ontologies. In other words, they are providing you with a kind of semantic indexing service, which enables other people to find your information faster and more easily. You might well be willing to pay for that.”

Q: “Will the ordinary web-surfer notice anything of this? From what you are telling me, it mostly seems a back-office operation. What will change for the everyday internaut?”

Van Harmelen: “The Semantic Web will be mainly a success when it remains invisible. All the technology we were talking about before is indeed "below the surface". The only thing that a web-surfer will notice is that the quality of the results he gets back from the search engine are a lot better then before. The current search engines are very good at recall: they find everything there is to find. But the score not so well on precision: besides finding what you were looking for, they give you a whole lot more that you didn't want. Of course, I'm exaggerating a bit, but the current precision should be improved a lot. And I would also expect improvements in the way the information is presented to the users. For example, if I currently type my name ‘Van Harmelen’ in a search engine, I get two types of results: some are about me and my scientific work, while others are about the Dutch village Harmelen. The problem is that the search engine cannot distinguish between the two, and simply puts all information in a single big list. When the Semantic Web will be in place, the search engine should be able to determine that there are really two types of hits, and that these should be displayed separately, or even better, it should ask me what I was looking for: the person Frank van Harmelen, or the village Harmelen.”

Q: “So searching will become easier. Any other advantages?”

The future of the search engine: "Search and you will find"

Van Harmelen: “An important theme that we haven't discussed is personalisation. If I look at a particular website, and you look at the same website, we both see the same information. Of course, that's far from ideal, since you have very different interests from me. Take again Amazon as an example: wouldn't they gain a lot if they could manage to show you a different page then me, tailored to your own interests? You could even use personalisation to reduce the information overload that we are all suffering from these days: everything that doesn't match your interest profiles doesn't need to be presented to you.”

Q: “How many people are currently working on the Semantic Web?”

Van Harmelen: “Actually, the W3C is a rather small organisation. They have a lot of members, but the size of their staff is very limited: worldwide only a few dozen people or so, and many of them are concerned with other things besides the Semantic Web. Employees of the member organisations are doing the real work in W3C. In the Semantic Web working groups you will find people from IBM, Hewlett-Packard, Sun, but also Nokia, Phillips and Daimler-Chrysler. You may be surprised at some of those names, but many different organisations stand to gain a lot from the Semantic Web. For example, Hewlett-Packard sees possibilities to use the Semantic Web to turn their printers into self-describing devices. Every printer will get its own profile, written in a Semantic Web language. And what happens? You walk into a building - say a conference centre - and all printers notify themselves to your laptop, or your PDA. And when you want to print something, your laptop or PDA has already decided which printer is closest to you, and most suited to your specific printing job. And a company like Nokia aims to make many new mobile services available through their phones. So you can understand why all these companies are joining the development of the Semantic Web. Nobody wants to be left behind.”

Q: “And what about Google, Alta Vista, Yahoo, etc? Are they involved in developing the Semantic Web?”

Van Harmelen: “I recently listened to some of the Google people, and I was surprised how, well, shall I say, politely reluctant they were. There were very well aware of the latest developments, but they told us that they were not actively exploiting the technology. But at the same time you can see that they are experimenting. Have you heard of the Open Directory? It's a project where thousands of volunteer classify countless web pages into tens of thousands of categories. Well, Google is already linking its search results with the Open Directory topic hierarchy. For many search results you will already find a hyperlink to the category where that result is classified in the Open Directory. That enables you to go and look at the other items in that category. So, without admitting it, they are already using Semantic Web technology. And of course, they cannot afford to ignore it. After all, the popularity of a search engine is only determined by the quality of its results.”

Q: “And what about other applications?”

Van Harmelen: “A concrete example is the ontology developed by W3C to describe device capabilities. That explains to a computer what a particular type of telephone can and cannot do, what a printer can and cannot do, etc, and what kinds of information can be exchanged by these devices. Large ontologies already exist for specific domains. For example, the biomedical sector has developed large and high quality ontology describing medical terms and drug names. The car industry is also quite advanced; Daimler-Chrysler is even member of the W3C working group. Of course, these are applications that are not visible to the ordinary web-users. The main application in the short term will be in the business-to-business markets.”

Q: “And when can we expect to see the first applications for ordinary surfers and consumers?”

Van Harmelen: “Currently, I see many Semantic Web "islands" that are being developed within specific sectors. In the long run, I would expect these islands to merge and then you will get a real Semantic "Web". Only then will things get interesting for the average consumer. Something that I see happening soon - and Philips is already quite active in this area - is the development of ontologies for multi-media contents. An example would be the web sites that listing the daily television programmes. Currently, those web sites are only readable for humans. But with metadata, based on an underlying ontology, my computer or my PDA could read such pages, compare the listings with my interest profiles, and alert me to interesting programmes that are coming up. In the same way, I could imagine ontologies for music styles, or film genres. All that you need to do then is to indicate which styles or genres you like, and the computer can do the rest. I would expect those applications to appear in the next few years.”

Q: “A difficult question to finish with: when do you expect the big breakthrough of the Semantic Web?”

Van Harmelen: (grinning) “That is indeed a hard question. Predicting the future is always hard, and particularly in the IT sector. And predicting events on the World Wide Web is completely impossible. Tim Berners-Lee - the father of the World Wide Web - uses the metaphor of a bob sledge: First you have to push hard to get going, but once it's going you have to jump in quickly before it runs of without you. Well, the Semantic Web is currently in the pushing stage. We have to talk to industry to convince them of the benefits. But I don't doubt that the Semantic Web will happen. Tim Berners-Lee also sees the Semantic Web as the only way forward for the World Wide Web. But let me be more specific: I would be very disappointed if in two or three years time we wouldn't see any applications for ordinary web-surfers, particularly in the e-commerce area. That sector has much to gain from personalisation. To reform the entire Web into a Semantic Web will take much more time. But I'm convinced it will happen”.