

Scientific research is changing

Martijn de Groot

The World Wide Web is functioning as a science accelerator, says full professor in Knowledge Representation and Reasoning Frank van Harmelen who is based in Amsterdam. That is changing things. But this change barely compares to what is yet to come.

‘Of course the web has changed our lives as scientists. As researchers, we now blog and hyperlink as if our lives depend on it,’ said the computer scientist in Amsterdam when he, quite appropriately, held the foundation day speech at the Free University (VU) on 20 October 2009 in the presence of honorary doctor and web inventor Tim Berners-Lee. He continued: ‘But that’s all same old same old. I argue that the web will have a much more profound impact on science in the near future than what we currently see around us.’

And Van Harmelen put it as follows at the jubilee symposium of DANS: ‘Whereas Web 1.0 was a network of images and text and Web 2.0 one of communities – by groups of people for groups of people – Web 3.0 must be the semantic web, the network of our data: By computers for computers

and hence useful to people.’

Particularly two consequences of the web’s existence will be important, Van Harmelen already explained in the accompanying issue of the quarterly e-data&research: In the first place, the manner of publishing and the possibility of exchanging data, and secondly, the possibility of using the web as a data source. ‘So far, publishing took place in scientific articles, which in fact is a sort of official farewell ceremony for scientific results. You create hypotheses, tables and plots, and no one else can use them afterward. No one can verify the data – because no one can access them. But neither can anyone use them for different, new purposes. That is all going to change. Now we can publish the data so that everyone can have them at their disposal. And right away, as opposed to a few years after

the researcher has wrapped up his or her work.’

Like a wave

This leads to the emergence of a plethora of new possibilities, argues Van Harmelen, and not just for continuing someone else’s work but notably also for combining data and hence answering completely new research questions. This way of publishing is already quite common in some of the natural and physical sciences, such as astronomy, particle physics and the life sciences. But it isn’t yet in a few other disciplines, like chemistry and the social sciences. ‘This movement is traversing science like a wave, however.’ It does require more than the mere possibility for exchange, emphasizes the computer scientist. ‘You have to do it in such a way that the data are usable to others. The semantic web has delivered a nice standard to that end: The Resource Description Framework – or RDF – which describes ‘things’ and the ‘relationships between things’. Data sets consist of objects and relationships between objects, and of

variables and relationships between variables. RDF lets you describe the objects, variables, and the relationships among them for every data set. You end up with a labeled network of data. That makes it easier to reuse them and particularly for having that done by computers. In the life sciences, this is already fairly normal. There, they convert their data to RDF format as soon as they start sharing. That makes it easier to reuse them.'

The web as observatory

The social sciences already have examples of projects that use this approach to arrive at promising results, according to Van Harmelen. Communication scientists, for example, have their computers carry out content analyses of media items – instead of using a small army of students or stay-at-home moms – and store the data in RDF so that they can easily link them to data from many other sources. This turns the web into an observatory, on top of it being an efficient medium for data exchange. The VU's Network Institute, an alliance that includes computer scientists, economists and social scientists, studies the effect of media on politics. It does that in conjunction with communication scientist Jan Klein Nijhuis. 'In the past, they would hire small armies of stay-at-home moms



Figure 1. The web itself is turned into an observatory, on top of it being an efficient medium for data exchange (internetcafé in Vietnam).

and students, literally, to collect data for a content analysis. Nowadays, a computer reads those data and stores them in RDF format, thus linking them to all sorts of other studies and analyses.'

The Network Institute is also looking into the scientific process, together with Professor Peter van den Besselaar of the Rathenau Institute. Besides citation analyses – 'that is not a very large data set and it lags five years behind,' says Van Harmelen – it uses the entire web. 'Scientists do a lot more than just quote and publish.

They also attend conferences and write on blogs, and as it's all visible on the web, you can measure it. In the long run, you no longer will have to hold researchers accountable on the basis of their publications alone, but there will be new methods for measuring the effect and the quality of their work.'

Data want to be free!

A similar approach – using the web as observatory – is applied in a study by organization scientists at the VU, into knowledge networks among

businesses. 'Is it possible to predict startups as a function of the way of participation in knowledge networks? In the past, you would send questionnaires for such a study, and you would be only too pleased if your response rate was 30%. Increasingly, those networks can be traced on the web. Just look at LinkedIn.' One condition is that as many data as possible are freely available. 'Data want to be free!'

That leads to the question as to how reliable the sample quality is. Old-fashioned questionnaires did allow you to determine the exact composition of the sample, but if you grab data off the web, you have to wait and see who did and who did not place theirs online. Van Harmelen: 'In the old days, you didn't know who was going to reply so the problem was essentially the same. And of course you can still send questionnaires. The point is that an entirely new observatory has been added!'

What to do

The developments are rapid and the perspectives are very attractive, Van Harmelen feels. But they're not only attractive. There is such a thing as going too fast, and he quotes part of the title of one of his publications: 'What to do when success is becoming a problem?'

This contribution is an edited version of a previous article by Martijn de Groot in the quarterly e-data&research

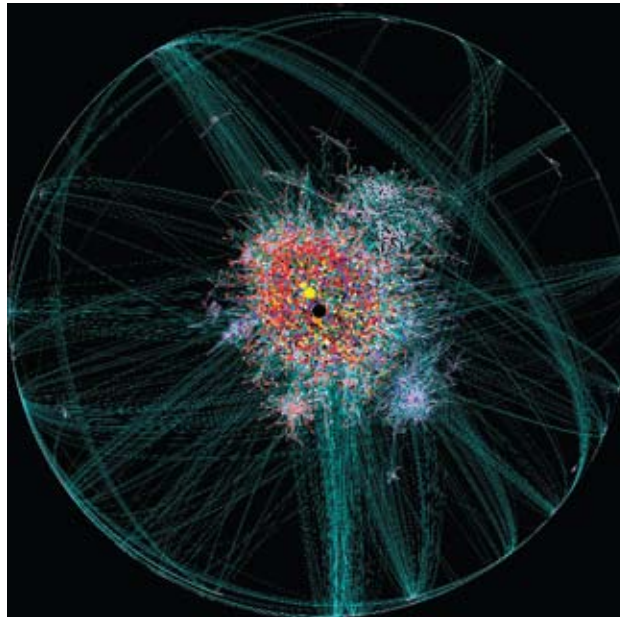


Figure 2. Computer generated map showing relationships between weblogs. This map shows blogs as coloured spots and lines represent the links between them. Nodes within the centre are of a more technical nature; nodes at the outer part of the sphere are non-technical blogs. Colours represent the URL. This image was created by Matthew Hurst.