Bus Inf Syst Eng 58(5):371–373 (2016)
DOI 10.1007/s12599-016-0448-y

PROFILE

Interview with Frank van Harmelen on "Linked Data and Business Information Systems"

Sören Auer

Published online: 24 August 2016

Prof. Frank van Harmelen
Knowledge Representation and Reasoning Group
Department of Computer Science
Faculty of Sciences
Vrije Universiteit Amsterdam,
de Boelelaan 1081a,
1081HV, Amsterdam,
The Netherlands

Frank van Harmelen (1960) is a professor in Knowledge Representation and Reasoning in the Computer Science department at the Vrije Universiteit Amsterdam. Since 2000, he has played a leading role in the development of the Semantic Web. He was co-PI on the first European Semantic Web project (OnToKnowledge, 1999), which laid the foundations for the Web Ontology Language OWL. He co-authored the Semantic Web Primer, the first academic textbook of the field and now in its third edition. He was one of the architects of Sesame, an RDF storage and retrieval engine, which is in wide academic and industrial use with over 200,000 downloads. This work received the 10-year impact award at the 11th International Semantic Web Conference in 2012.

In recent years, he pioneered the development of large scale reasoning engines. He was scientific director of the 10 m euro EU-funded Large Knowledge Collider, a platform for distributed computation over semantic graphs with billions of edges which has improved the state of the art by two orders of magnitude.

He is scientific director of The Network Institute. He is a fellow of the European AI Society ECCAI, member of the Academia Europaea and the Royal Netherlands Society of Sciences and Humanities. He is a guest professor at the University of Science and Technology in Wuhan, China.

BISE: Just as there was a software crisis in the early days of computer science, do you think we are now in the middle of a data crisis?

Van Harmelen: On the contrary! The software crisis was about costs that were spiraling out of control, and the quality that was poor (in those days: of software). For data, we are now not seeing a crisis but a boom: costs are dropping dramatically (both the costs of data and the costs of the analytics) and both the volume of data and the quality of the analytics are constantly increasing.

BISE: How do you see the interplay of Linked Data with information systems and business?

Van Harmelen: Linked Data is crucial as soon as you start to combine multiple data sources. When I talk to business people, I tell them: if you are using 10 data sources, you have 45 integration problems (the number of links between 10 data sources), and if you add only 5 data sources, your problems suddenly more than double, to

Prof. Dr. S. Auer (✉)
Computer Science Department/Enterprise Information Systems,
University of Bonn, Römerstr. 164, 53117 Bonn, Germany
e-mail: auer@cs.uni-bonn.de

more than 100 integration problems. In other words: in a world where we are increasingly combining multiple heterogeneous data sets from multiple sources, Linked Data technologies are crucial.

BISE: What are the main advantages of RDF, Linked Data and semantic technologies compared to other data integration techniques such as Web Services, XML or data warehousing?

Van Harmelen: The key difference is that semantic technologies try to make as much of the intended meaning of the data explicit in a computer processable way. The more our software has access to the intended meaning of data, the cheaper it will be to integrate heterogeneous datasets. Let me give a simple example. In an XML document, if the tag <capital> is nested inside a tag <country>:

```
<country>The Netherlands
      <capital>Amsterdam</capital>
</country>
```

What does that mean? What is the intended relationship between <capital> and <country>? Is a <capital> part of a <country>? Is a <capital> a type of <country>? Is a <capital> located in a <country>? We all know it is the latter, but the XML does not tell us that. A semantic model would tell us that the intended meaning of the dataset is "located in", together with some of the consequences that this entails. For example: a capital can only be located in 1 country, so if Amsterdam is located in The Netherlands as well as in Les Pays-Bas, then apparently those two names must designate the same country.

In short: *adding explicit semantics reduces the scope for misunderstandings* when you are re-using somebody else's dataset, in particular when you are trying to integrate multiple datasets that you did not create yourself.

BISE: What are the main research challenges in the Linked Data area?

Van Harmelen: The good news is that after 15 years of R&D, the main building blocks are now in place. We understand how to publish data on the web in such a way that it becomes possible to make links between data sets, to integrate multiple datasets, and to find them and reuse them. This is also apparent from the rapidly increasing number of large and reputable organizations that are now adopting Linked Data principles in their data management processes: governments (UK, US, EU), media (BBC, New York Times), retail (Kmart, Sears, BestBuy), industry (XMP, Volkswagen, Renault), the cultural sector (e.g. the Prado in Spain, the Europeana library in the EU, the Getty Foundation in the US), and search engines (Google, Yahoo, Bing).

However, some elements that are crucial for continued and wider adoption: trust and data quality are crucial. *Provenance* is an important mechanism for this, but often not yet used. Many applications are dealing with highly volatile data, and we are now seeing an increasing interest in dealing with *streaming* semantic data. Other challenges are partly technical, partly social: good *pay-models* for Linked Data (since not everything on the web needs to be for free) and good *privacy-protection* mechanisms (since not everything on the web needs to be open).

BISE: How should the balance be between lightweight semantics and full-fledged knowledge representation facilitating reasoning?

Van Harmelen: We are seeing different applications that cover different parts of this spectrum. By far the most applications use only a very lightweight semantic model, with a type-hierarchy, and perhaps some mapping between datasets using owl:sameAs links. But something like the widely used SNOMED ontology, describing hundreds of thousands of terms used in medical practice, has been heavily formalized using OWL.

Also, we are seeing that the more expressive forms of reasoning are used "off-line", when developing a vocabulary, or when constructing mappings between datasets, while the "on-line" use of reasoning is often limited to simple traversals of type-hierarchies for tasks such as product recommendation or query expansion.

So, in short, there is a role for both, depending on the type of application, and depending on their use in the development process.

BISE: Is it reasonable to add a data interoperability layer in enterprise architecture and integration?

Van Harmelen: The question is no longer whether such a data-interoperability layer is reasonable, since it is increasingly absolutely necessary. Increasingly, enterprises are integrating multiple datasets, both from different sources inside the company and from sources outside the company. The prize-winning work of the semiconductor company NXP is a very nice showcase for this: using semantic technologies to integrate information about more than 10.000 products from dozens of different databases from different parts of the company. See their whitepaper at http://ceur-ws.org/Vol-1383/paper3.pdf and their slide deck at http://www.slideshare.net/parvathymeenakshy/applying-semantic-web-technologies-1. Their entire product catalogue (at http://www.nxp.com/products/products:PCPRODCAT) is now run from a semantic model described in RDF and SKOS. And they are or course not the only one. We know of similar stories from many of the different players I discuss elsewhere in this interview.

BISE: Is the concept of enterprise knowledge graphs a good candidate to realize such an interoperability layer and what steps would be required to further refine and consolidate this concept?

Van Harmelen: Indeed, the term "enterprise-wide knowledge graph" is increasingly being used. Yahoo was one of the first in building a company-wide knowledge graph, and they used to drive advertisement recommendations.

Similarly, Elsevier Publishing, one of the largest science publishers in the world, is transforming itself into a digital data and services company, and as part of this they are building very large company wide knowledge graphs that aggregates information about journals, authors, papers, topics and citations. They do this to offer more integrated data-services to their clients, both universities and large knowledge-intensive industries such as the pharmaceutical industry.

A final example is Deloitte, who are building a large knowledge graph containing information on companies, their financial and ownership data, their patent portfolios, market indicators, tax regimes in different countries, etc., etc. They heavily invest in building this graph in order to support the Deloitte consultants when they give advice to their clients.

All of these large companies have in recent years scooped people away from my lab for bigger salaries than I could offer in order to acquire expertise on semantic technologies and knowledge graphs.

BISE: In which application areas is Linked Data already well established and which are yet to be developed from your point of view?

Van Harmelen: The most successful areas are already quite broad: government, media, manufacturing, online retail, cultural heritage and science. The two most promising new areas of application are in my opinion healthcare and the internet of things.

The healthcare sector is suffering from extreme fragmentation of data, and from very poor integration between data sources. Every year we see press reports on unnecessary medical incidents due to poor data-transfer between healthcare organizations, or even between different departments inside a single hospital. Better data-integration will lead to reduced costs and increased quality, and semantic technologies make a crucial contribution to this data-integration problem in the healthcare sector. I could do an entire interview only on this, there are so many opportunities there. These problems exist in every advanced healthcare system, both in the US and in the EU. As an early success, Kaiser Permanente has reported $1 billion in savings from reduced office visits and lab tests as a result of ensuring data exchange across all medical facilities.

An area with which I am less familiar, but which I hear is rapidly gaining in importance is the Internet of Things. Increasingly, the devices we use in our daily lives are becoming interconnected, including devices that have until now no connectivity at all. The obvious examples come from the home environment (heating, lighting, kitchen appliances), but increasingly this also concerns the urban environment, from street lights to garbage containers. Semantic technologies and Linked Data can play a crucial role in helping to integrate the data streams that come from such a highly connected environment. There is much activity on this in the Far East, in particular China and South Korea. Europe should become much more active in this space, including the use of Linked Data to facilitate this increased connectivity.