

# How will we interact with the Web of Data?

Tom Heath, Talis Information Ltd

This is a pre-print (minus layout and a few editorial tweaks) of the article:

*Tom Heath (2008) How Will We Interact with the Web of Data? IEEE Internet Computing, Vol. 12(5), pp. 88-91.*

The Semantic Web is a global information space of linked data, designed not for human use but for consumption by machines. Right? Well, yes and no. It's true to say that machine-readable data, given explicit semantics and published online, coupled with the ability to link data in distributed data sets are the key selling points of the Semantic Web. Together, these features allow aggregation and integration of heterogeneous data on an unprecedented scale, and machines will do the grunt work for us.

However, without a human being somewhere in this process, to reap the rewards of these new capabilities, the endeavour is meaningless. Far from removing human beings from the equation, a Web of machine-readable data creates significant challenges and significant opportunities for human-computer interaction.

To date, the Semantic Web community has mostly been busy developing the technical infrastructure to make the Web of Data feasible in principle and on publishing linked data sets in order to make it a reality. If we are to fully exploit the challenges and opportunities of a Web of Data from a human perspective, we need to move beyond the initial phase and work to understand how this changes the user interaction paradigm of the Web.

In this column I'll discuss some ways in which our interaction with the Web of Data may differ from how we interact with the established Web of Documents, and what this might mean for both users and producers of content on the Web.

## ***Semantic Web: from vision to reality***

In 1999 Jakob Nielsen wrote about a looming crisis (Nielsen, 1999). The Web was growing at a phenomenal rate, and without closer attention to user interface principles he predicted the Web would become an unusable mass of documents. Almost ten years later, the Web is undergoing another seismic shift. The result of this shift is the emergence of the Web of Data, or Semantic Web; envisioned for more than a decade, and the result of many years work on the underlying technology. Whilst we may refer to them as distinct concepts, the Web of Data is not a separate entity removed from the Web of Documents, but more akin to another layer of cloth interwoven with the Web as we know it.

This time around, in 2008, the headline statistics for the growth of the Web are not quoted in terms of Web pages or Web sites. Instead people talk about numbers of 'triples' published on the Web of Data using the 'Resource Description Framework' (RDF), and the number of links these triples create between distributed data sets.

RDF is a W3C specification for making statements about things in machine-readable form. These statements each consist of a 'subject', 'predicate' and 'object', hence the name 'triples'. In most cases, the subject of a triple is a Uniform Resource Identifier (URI) that can identify anything the

data publisher chooses, be that a person, a place, a document on the Web, an abstract concept – anything. Predicates specify the nature of the relationship between the subject and object, are drawn from vocabularies published on the Web and are also identified by URIs. The object of an RDF triple is usually a string literal or another URI. Where the object is a URI from a different namespace, i.e. it identifies something in an external data set, then the RDF triple creates a link between those data sets, replacing isolated 'data islands' with a giant, distributed data set built on top of the Web architecture – a true Web of Data.

When members of the grassroots Linking Open Data project last tried to calculate the current size of the Web of Data, their conservative estimates suggested that data sets in the Web contained more than 2 billion RDF triples, 3 million of which are links across data sets (Bizer, Heath et al., 2008). The rate of growth in this Web is so great that any future estimates are likely to be out of date as soon as they are published.

One other feature of RDF is worth noting at this point. RDF allows the easy integration of triples contained in any number of documents distributed across the Web. Source documents can be merged painlessly, without the graph that results from this merge needing to conform to a particular schema. One consequence of this is a major reduction in the headaches associated with integrating heterogeneous data.

## ***Throw out your homepage!***

On the Web of Documents great care is taken to develop visually attractive homepages that send out just the right message about the entity they represent. If RDF enables data from multiple sources to be easily integrated to form a coherent 'view' of a particular 'thing', what does this mean for how we publish data on the Web? It means that the Web page as we know it is dead.

Developers of Web2.0 mashups have been demonstrating this for some time, integrating data from a handful of different sources to present a novel view that none of the source data sets alone is able to provide. The Web of Data is the logical extension of this process, allowing developers to create links between data sources that are themselves exposed on the Web for others to reuse to build large-scale, ad-hoc mashups, whilst simultaneously reducing the headache involved in integrating heterogeneous data.

Documents will always be useful containers for data, but in many cases I predict they will become nothing more than that. On the Semantic Web you can't assume you have control over how the information you publish will be presented – it's just data. Thinking at the visual design level, RDF represents an extension of the long-established principle of separating content from presentation. For some this may ring alarm bells – how will brand be maintained if one has less control over presentation? For others it will represent an opportunity to free themselves from visual design concerns, concentrate in the first instance on publishing relevant, high-quality data, and let others build the views they want rather than those that someone else assumes they need.

At the data level, publishers can have some influence over which external sources their data is linked to, primarily by creating these links themselves and publishing them for others to consume. However, on the Web of Data one cannot control with any degree of certainty the sources with which one's data is integrated – enabling serendipitous reuse is exactly the point! As already discussed, data published on the Web in a reusable form enables new views that have value beyond the sum of the parts and that might not have been anticipated in advance.

It is for these reasons that I suggest we throw away our homepages. Researchers well know the challenge of connecting all the pieces of their professional activities into a coherent whole: the projects, the papers, the committee and editorial board memberships, the blog entries and photo

albums, all scattered across isolated islands on the Web, maybe replicated on their personal Web site or connected by strands of hypertext; or maybe not, due to the effort involved.

A homepage for the Web of Data takes a different shape. At the most basic level it may simply be a collection of RDF triples that tie together data we want to express about ourselves that is distributed across numerous locations. The job of the machines is then to assemble this data into a coherent view, ready for human consumption.

To put my money where my mouth is, next time I get business cards printed I won't be including the address of my homepage. Instead I'll put my URI on the card, safe in the knowledge that a human being with a browser, Semantic or otherwise, can look up that URI and find some of what the Web has to say about me.

## ***What should a Semantic Web browser look like?***

Extending these ideas, we can see that the document in which a particular RDF graph is published becomes primarily an indicator of provenance, rather than representing the definitive packaging of a certain slice of data or content.

Of far greater relevance than the documents themselves are the things described in those documents – the people, places, concepts, etc. So far I've talked about a Web of Data, but when I do this I'm really using it as shorthand for 'Web of Data about Things' – any things. One might not be able to retrieve a car over HTTP, but one can identify it with an HTTP URI and use the Web to retrieve its description in RDF.

It is at the level of 'things' that browsers for the Web of Data should operate. Providing simple browsers for RDF triples, and the documents in which they are published, is one option for enabling people to interact with this information space. We have seen this trend with some of the earliest Semantic Web browsers, but it rather misses the point. The one-page-at-a-time style of browsing, that we know well from the Document Web, would make nothing of the potential we now have for integrated views of data assembled from numerous different locations.

Therefore, Semantic Web browsers must not simply echo the underlying representation of the data (Karger and schraefel, 2006) by presenting a view on individual documents that contain RDF triples. Instead they must treat 'things', in the broadest sense, as first class citizens of the interface. The particular thing of interest should take centre stage, with the browser assembling relevant information seamlessly behind the scenes.

We are seeing shades of this trend in Semantic Web browsers such as The Tabulator (Berners-Lee, Chen et al., 2006) and DBpedia Mobile (Becker and Bizer, 2008), where the thing of interest is of greater importance, and specific documents simply supply fragments of data that together make up a broader picture. Despite these moves in the right direction, there is some way to go yet.

Conventional browsers have largely failed to deliver on the original vision of the Web as a read/write medium. Whilst this vision is slowly being realised at a general level through, for example, blogs, wikis and specialised annotation interfaces such as Flickr, there remains a significant degree of indirection when it comes to editing Web documents. In some cases this process still involves starting an editor for HTML documents, making appropriate changes and then starting some other application (such as an FTP client) in order to publish the updated document.

Browsers for the Semantic Web, 'thing browsers', have an opportunity to enable a far greater degree of direct manipulation in their interfaces. Different types of objects afford different types of actions, and knowing the type of object on which the user is focused should allow browsers to

provide menu of actions that are specialised for this type of object, and perhaps even adapt these according to the context.

For example, if the user is currently browsing a person, the browser may enable the user to send a message to that person, share an object with them, or arrange a meeting, without any of these functions having been explicitly listed as actions that can be invoked on these individuals. Instead, the Semantic Web at large may provide the necessary knowledge and services on which to offer such functionality, such as statements describing 'arrange meeting with' as a valid action for a thing of type 'person', or definitions of what constitutes a meeting, or venue suggestions that are tailored to the relationship between the two parties and the time of day.

Clearly a Web of Data is unable to offer direct manipulation of 'real-world' things, such as cars and dogs, which are not, and never will be, online. However, in a Web where we can explicitly reference anything, not just documents, there is great potential to reduce the degree of indirection in Web interfaces. We no longer have to refer to Web pages about things, but can refer to the things themselves.

In case there was any doubt, this is no overnight endeavour but a trend that will take many years to be realised and may take many different forms. Giving a keynote talk at WWW2007, Bill Buxton made the claim that "*The diversity of 'web browsers' tomorrow will match the diversity of 'ink browsers' (a.k.a. paper) today - in terms of diversity of form, function, location, and importance*". I don't get the impression that he was thinking about the Web of Data when making this statement, but the claim stands up nonetheless – the diversity in a true Web of Things will require similar diversity in the interfaces through which we exploit it. The browser is just one approach.

## ***A Back button for the Semantic Web?***

Accepting the shift from document to thing, and from predefined views to those assembled dynamically, will not just require completely new interfaces, but also a number of changes to the interaction widgets in interfaces with which we're already familiar. If browsing becomes not just about moving from one document to another by following links, but about integrated views of data assembled from a variety of sources, then the notion of the 'Back' button takes on a slightly different meaning in the interface. Rather than moving between documents, the Back button in a Semantic Web browser should move the user to previously viewed things. More significantly, a form of 'Undo' button, as you might find in a word processor, could be of critical importance in an environment where vast amounts of data can be assembled at minimal cost, but not all of it will be pertinent to the job in hand.

The range of potential sources from which data will be available about a certain thing will be immense. Imagine entering a URI for "London" into the address bar of your Semantic Web browser. All the data available on the Web about this thing can not feasibly be presented in one interface; users will need to decide which sources to add in depending on their current task or context, or will need this decision to be made intelligently for them, with the ability to undo the addition of any particular sources. This functionality becomes even more critical if automated reasoning is carried out on Semantic Web data, creating knowledge which was not previously explicit in any of the individual data sources.

How to manage the assembly of these data sources becomes a critical issue. When a number of colleagues and I evaluated the deployment of various Semantic Web technologies to delegates at the 2006 European Semantic Web Conference, one of the key themes to emerge was 'coherence' (Heath, Domingue et al., 2006). Delegates had been presented with various Semantic Web applications for use at the conference; they expected data to be integrated across these and presented as a coherent whole. For various reasons described in [ref] this was not possible, leading to a suboptimal user experience and confusion for delegates.

Key to developing Web of Data browsers will be look up services such as Sindice (Tummarello, Oren et al., 2007), that provide a means to find other RDF documents on the Semantic Web that mention a particular thing. This kind of service may help ensure that the user experience is coherent, in that it includes all data the user would expect it to. However, ensuring that a particular view of data is useful is another question.

Any system aiming to integrate heterogeneous data on an ad-hoc basis, and present this to users, will need to adopt sophisticated models of relevance, quality and trust that are sensitive to the current task and context of the user. How that might be achieved is a question for another day.

## **References**

Becker, C. and Bizer, C. (2008) DBpedia Mobile: A Location-Enabled Linked Data Browser. In Proceedings of the Workshop on Linked Data on the Web, 17th International World Wide Web Conference, Beijing, China.

Berners-Lee, T., Chen, Y., Chilton, L. et al (2006) Tabulator: Exploring and Analyzing linked data on the Semantic Web. In Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI2006), 5th International Semantic Web Conference, Athens, GA, USA.

Bizer, C., Heath, T., Idehen, K. and Berners-Lee, T. (2008) Linked Data on the Web (LDOW2008). In Proceedings of the 17th International World Wide Web Conference (WWW2008), Beijing, China, April 2008.

Heath, T., Domingue, J., and Shabajee, P. (2006) User Interaction and Uptake Challenges to Successfully Deploying Semantic Web Technologies. In Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI2006), 5th International Semantic Web Conference, Athens, GA, USA.

Karger, D. and schraefel, m.c. (2006) The Pathetic Fallacy of RDF. In Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI2006), 5th International Semantic Web Conference, Athens, GA, USA.

Nielsen, J. (1999) User interface directions for the Web. Communications of the ACM, Vol. 42, Issue 1, pp. 65-72.

Tummarello, G., Oren, E., and Delbru, R. (2007) Sindice.com: Weaving the Open Linked Data. In Proceedings 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, Busan, Korea.