# Nodalities

THE MAGAZINE OF THE SEMANTIC WEB

www.talis.com/nodalities

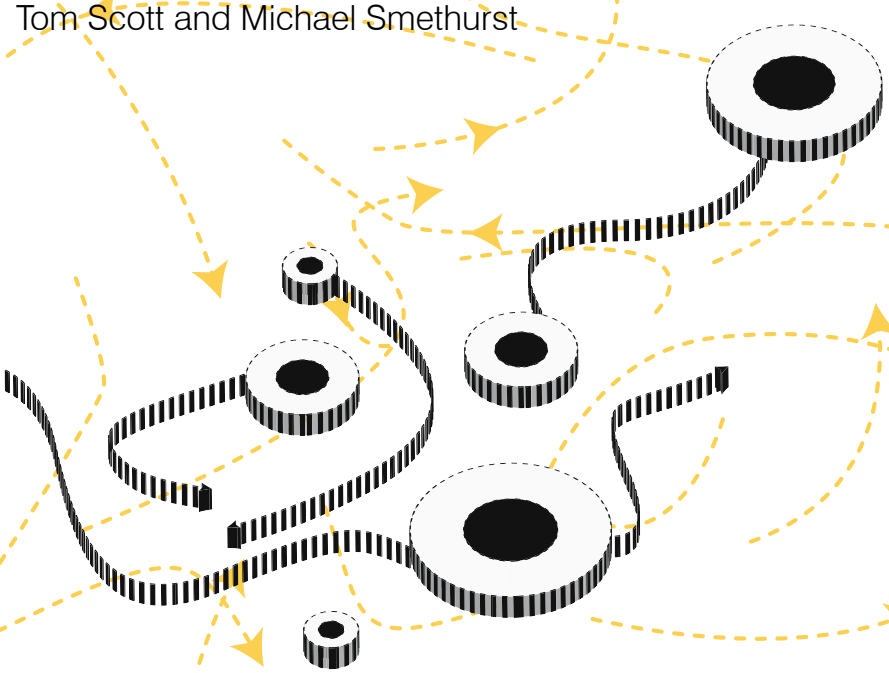
November / December 2008


INSIDE



# Building coherence at bbc.co.uk


Tom Scott and Michael Smethurst

## Telling (non-linear) stories

For the past 86 years, the BBC has plied its trade as a storytelling organisation. In the world of linear broadcasting we've even gotten very good at it. Guiding the audience through complex news story lines, explaining the natural world and, interleaved narrative arcs and the plotlines of drama has become our forte. But storytelling in a linear world is different from storytelling in the non-linear, hypertext world of the web.

With the exception of BBC News Online (news.bbc.co.uk) the online world has often been seen as a supporting adjunct to the linear broadcast world. Over the years we've commissioned and built sites to provide online support for programmes; but we've too often taken our linear storytelling expertise and attempted to replicate the same techniques on the web - with mixed success. Unlike linear broadcast storylines the web doesn't provide


*continued on page 3*

## SUBSCRIBE NOW

Nodalities Magazine is made available, free of charge, in print and online. If you wish to subscribe, please visit www.talis.com/nodalities or email nodalities-magazine@talis.com.

## MEET US AT

Talis staff will be presenting and attending several events over the next few months, including;

**IUI2009**
Sanibel Island, Florida | 8-11 February 2009

**ESWC2009**
Heraklion, Greece | 31 May - 4 June 2009

**WWW2009**
Madrid, Spain | 20-24 April 2009

**For further information visit: www.talis.com/platform/events**

**Nodalities Magazine is a Talis Publication**

ISSN 1757-2592

Talis Group Limited
Knights Court, Solihull Parkway
Birmingham Business Park B37 7YB
United Kingdom
Telephone: +44 (0)870 400 5000

**www.talis.com/platform**
**nodalities-magazine@talis.com**



## EDITOR'S NOTES

Welcome to this fifth edition of Nodalities Magazine. 2008 has seen progress from the Semantic Web space through the launch of applications and startups, and through many conferences and events. If you haven't read any of the previous editions, they can be found on www.talis.com/nodalities. From there, you can also sign up to receive future Nodalities in print free of charge. If anything in Nodalities Magazine has sparked your interest, feel free to join the discussion on our blog from the same link above.

This issue features Tom Scott from the BBC telling non-linear stories about surfacing BBC programme information as Linked Data. The BBC produces huge amounts of content, and have a well-established online presence. However, as Tom Scott explains, they are trying to weave this content more intimately with the wider web, and their answer has been Linked Data.

Elsewhere in this issue, Bill Roberts introduces us to the story of semantic wikis with Swirrl. From Sweden, Anders Söderbäck and Martin Malmsten have introduced the Semantic Web to the world of Libraries, where the LIBRIS library catalogue relies heavily on semantic web technologies. They tell how they came to choose RDF as a means of organising these huge literary datasets. We also hear from Leigh Dodds, programme manager at Talis, who discusses the concept of the Linked Data ecosystem, and explores how Talis is working towards enabling it.

We also have an article by Justin Leavesley about the world becoming more heavily connected: from desktops to washing machines, the future is online. Tom Heath also looks back at the International Semantic Web Conference (ISWC2008)—a conference which has "come of age" this past year. Finally, we feature a podcast transcript of David Provost talking with Talis about his Semantic Web industry report "On the Cusp".

If, while reading, you feel you'd like to contribute an article or have a suggestion, please feel free to get in touch with me at zach.beauvais@talis.com. It will be an interesting 2009, and I'm very much looking forward to the innovative creativity I've witnessed throughout this year continuing and facing up to the challenges of a new year.

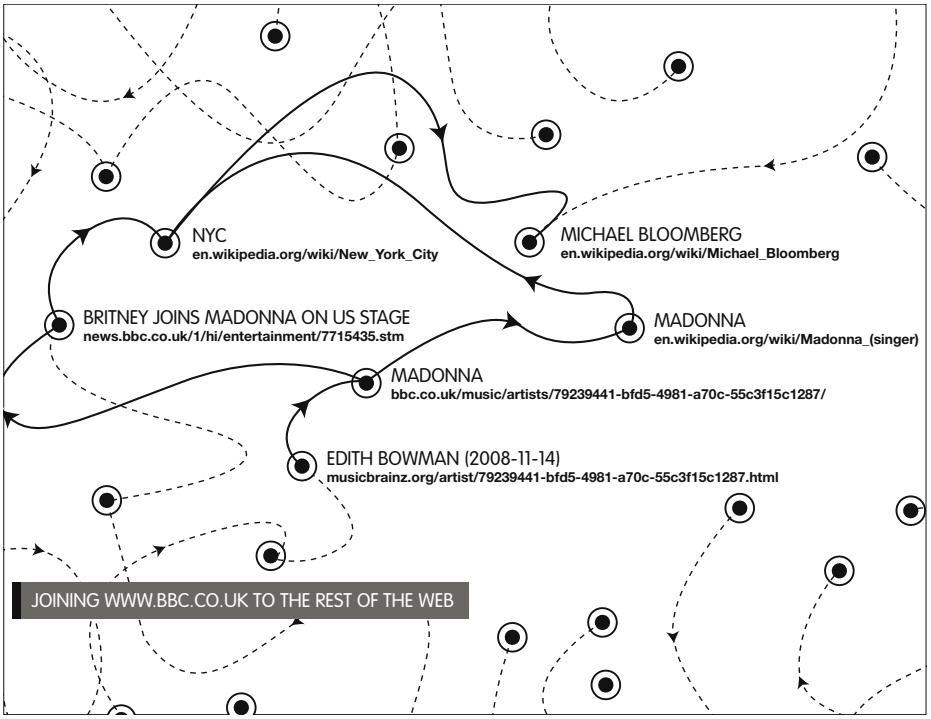
*Continued from front page.*

people with a predicted and controlled linear journey. Instead we dip in and out of any given website - following different journeys - to find the information we want at that time.

Many of our programme support sites have been commissioned and developed in isolation. So you see an Archers site and an Eastenders site and a Top Gear site which are internally coherent but which fail to link up other than via editorially determined cross promotions. Want to see who presents Top Gear? No problem, we can do that. Want to see what else those people present? Sorry, can't do that. By developing self-contained microsites the BBC has produced some good stuff but it has also been unable to reach its full potential because it hasn't managed to join up all of its resources. By failing to link up the content (on both a data and a user experience level) the stuff we publish can never becomes greater than the sum of its parts. Without these links we can't make bbc.co.uk a coherent experience. As a user, it's very difficult to find everything the BBC has published about any given subject, nor can you easily navigate across BBC domains following a particular semantic thread. For example, you can't yet navigate from a page about a musician to a page with all the programmes that have played that artist.

> *We had to recognise that non-linear storytelling puts the narrative arc into the hands of the user*

So how do you tell stories on a web scale? We could stick with the easy option and try to control 'user journeys' across the site. Provide links to where we think the user should go next. But that's little better than those flip a dice, go to page 30 dungeons and dragons books we all had as kids. We had to recognise that non-linear storytelling puts the narrative arc into the hands of the user. What to read, what to click, where to go next is really up to you. So storylines split and merge, meta-narratives emerge and fracture; 'user journeys' slip out of (editorial) control.




All of this comes from the power of the link - back to basics. But we can only provide precisely targeted links at the user experience level if those links exist at a data level. And that's the difficult part. The organic growth of our sites has been mirrored in the organic growth of our content and data management systems. We currently have a range of systems across the business for managing different bits of content throughout the production chain. And like our public facing sites none of these speak the same language or share the same identifiers. A typical episode of Top Gear might have 6 separate identifiers on it's way from scriptwriter to airwaves to archive. Once you've solved this problem you hit the problem of multiple identifiers for James May and once you've got one canonical James May you're back to the problem of multiple identifiers for all the other programmes he's presented...

Solving these problems makes for a more linked, more coherent bbc.co.uk. But an internally coherent bbc.co.uk isn't enough. bbc.co.uk needs to be weaved into the rest of the web, not merely on the web. It needs to be linked in to all those other Top Gear / James May pages out there... Luckily the tips, tricks and techniques pioneered by the Linked Data community give us some clues here.

Add into this mix the fact that there's some data the BBC can never hope to provide. So we know when an artist is played on radio or TV. But we can't hope to know when they were born, or where they were born, or which bands they've been in, or who they're married to etc. If we want to tell stories around music all this is important data. And we can only get it by tapping into the collective knowledge of the web.

## BBC in the web of data

I'd like to claim that when we set out to develop /programmes we had the warm embrace of the semantic web in mind. But that would be a lie. We were however building on very similar philosophical foundations.

In the work leading up to bbc.co.uk/programmes we were all too aware of the

importance of persistent web identifiers, permanent URIs and the importance of links as a way to build meaning. To achieve all this we broke with BBC tradition by designing from the domain model up rather than the interface down. The domain model provided us with a set of objects (brands, series, episodes, versions, ondemands, broadcasts etc) and their sometimes tangled interrelationships.

We were also convinced that the value in programme websites lay not in the implicit metadata of the domain model but rather in the way this domain model overlapped and intersected with other domains. As ever the links are more important than the nodes because that's where the context lives: programmes:segment <features> music:track, programmes:segment <features> food:recipe etc. In this way we could weave new 'user journeys' into and out of /programmes, into and out of bbc.co.uk. From archive episodes no longer available online, to a recipe page, to a chef, to another recipe and back to a recent episode. Using well targeted content specific links we could not only escape the dead end content silos that characterised bbc.co.uk but point users back to programmes that would hopefully inform, educate and of course entertain.

Finally we believed in the merits of opening our data and building on top of other people's open data. When we looked to rebuild bbc.co.uk/music we looked at a number of commercial providers of music metadata. They all did a similar job to MusicBrainz (musicbrainz.org) - similar models, similar data quality etc. But choosing to go with a commercial provider would have precluded our ability to provide any kind of machine friendly (API if you must) views. The decision to publish JSON or vanilla XML or RDF would have been a decision to give the 3rd party business model away. So we went with the open alternative - an open, public domain provider, one that is more in keeping with our public service remit and one that represents better value for money for the license fee payer - which has to be a lesson to someone.

*To achieve all this we broke with BBC tradition by designing from the domain model up rather than the interface down.*

Without ever explicitly talking RDF we'd built a site that complied with Tim Berners-Lee's four principles for Linked Data:

- Use URIs as names for things. - CHECK
- Use HTTP URIs so that people can look up those names. - CHECK
- When someone looks up a URI, provide useful information. - Well, if we're only talking HTML, RSS, ATOM, JSON etc. CHECK
- Include links to other URIs. so that they can discover more things. - Again if we're talking HTML only CHECK

By keeping everything in its right place we'd also built a sane, maintainable, scalable, accessible site that search engines love and could be easily evolved to add new features and functionality. So to anyone considering how best to build websites we'd recommend you throw out the Photoshop and embrace Domain Driven Design and the Linked Data approach every time. Even if you never intend to publish RDF it just works.

Around this time we met by chance with some people from the Linking Open Data community and the two worlds collided. Obviously TBL wasn't talking only HTML in the last 2 principles but aside from that the parallels were striking. We set about converting our programmes domain model into an RDF ontology which we've since published under a Creative Commons License ( HYPERLINK "http://www.bbc.co.uk/ontologies/programmes/" www.bbc.co.uk/ontologies/programmes/). Which took one person about a week. The trick here isn't the RDF mapping - it's having a well thought through and well expressed domain model. And if you're serious about building web sites that's something you need anyway. Using this ontology we began to add RDF views to /programmes (e.g. HYPERLINK "http://www.bbc.co.uk/programmes/b00f91wz.rdf" www.bbc.co.uk/programmes/b00f91wz.rdf). Again the work needed was minimal.

*Finally we believed in the merits of opening our data and building on top of other people's open data.*

So for those considering the Linked Data approach we'd say that 95% of the work is work you should be doing just to build for the (non-semantic) web. Get the fundamentals right and the leap to the Semantic Web is really more of a hop.

## Why bother with RDF?

For all the pages we've published we've only had a limited success at making this information available for others to use, to hack with and to build new services with. While we've not done a very good job of making bbc.co.uk a coherent experience for people the situation is worse for machines.

It is our belief that rather than publishing proprietary APIs it is better to use the ubiquitous technologies of URIs and

HTTP. This approach supports the generative nature of the Web, making it easy for third parties to build with BBC metadata without learning BBC specific APIs and at the same time providing the BBC and its users with immediate benefits.

Services like Flickr, Twitter and the like have in many, many ways followed the same principles we adopted for programmes and music -- or if they didn't then the end results look pretty similar -- they are wonderful services. However, if as a third party developer you want to deal with the semantics, accessing the data via the Giant Global Graph to find everything about a certain person, place or topic and you wanted to include data from Flickr then you will need to deal with the specifics of Flickr. I suspect that it wouldn't be that difficult for Flickr to add RDF representations - if they did then Flickr content would be part of a common way of doing things. We want BBC data to be part of a common way of doing things.

Our hope in making BBC data available as RDF is that we will make it as generative as possible - helping others to do interesting things with our data. The BBC has a public service remit, a remit that means it should look beyond its internal business needs to help create public value around useful technologies and around its content for others to benefit from. The longer term aim of this work is to not only expose BBC data but to ensure that it is contextually linked to the wider web. We have started along this path by linking to Wikipedia (DBpedia in the RDF view) and MusicBrainz from the artist pages but this could be extended for programmes and events.

### The Semantic Web Gang

**The Semantic Web Gang is a monthly round-table podcast hosted by Paul Miller and featuring a regular panel of commentators on the Semantic Web.**

**semanticgang.talis.com/**

# Doing more with your data

*Bill Roberts introduces Swirrl, and explains where the Semantic Web fits in with the world of wiki.*

**By Bill Roberts, co-founder of Swirrl**

Helping people to do more with their data is what we are trying to achieve with Swirrl. The 'data silo' problem is well-known and significant. We think we can help to link the silos together.

I spent many years as an IT consultant and through working with a large number of organisations I saw how many inefficiencies there are in the sharing and re-use of data. This wasn't because they were bad companies: on the contrary, many of them were large, respected, successful and forward-looking. But this is a hard problem and the bigger the group and the longer its history, then the harder it is.

Also there are trends in the way we work that are highlighting these difficulties: people are more mobile, working from home or travelling; companies change ownership frequently, meaning any large organisation is in a permanent state of trying to restructure and integrate with its latest acquisition; through partnerships, outsourcing, contractors and consultants the boundaries of companies are more fluid, meaning that sharing information across the firewall is more important.

On the other hand, the world wide web shows the potential of connecting up large quantities of information. The "Enterprise 2.0" concept, of bringing ideas from web collaboration and social media into the world of work, has the potential to add enormous value and it's a movement that seems to be building in momentum.

So what do people want to do with data, and how are we going to help them do it? They want to use it to answer specific and sometimes complex questions. Let's list some of the requirements:

1. You have to be able to find relevant data.

2. You have to be able to understand it.

3. You want to explore and analyse it.

4. Once you've found your answer, it would be nice to make it easier for the next person with a similar question.

The first step to finding data is storing it somewhere with shared access. Sounds simple: but a huge amount of information in most organisations doesn't get over this first hurdle. If you are working in a group where people are in different places or belong to different companies, then the problem is that much greater. So our first task was to provide an easy-to-use shared environment for storing and accessing information.

There are lots of ways of doing this: but we are big fans of the wiki approach, for its simplicity, and because it follows the web based model that many people are now familiar with, allowing you to build up a network of links between related items. With a hosted wiki like Swirrl, the information is accessible anywhere with an internet connection and a web browser. You don't need to worry about installing client software, or worrying about client compatibility problems (OK, as any web developer knows, there are still compatibility issues across different browsers, but it's a problem we can live with.)

It has to be as easy to use as creating a spreadsheet on your own hard drive, because if it's significantly more difficult than that, then people will fall back into bad old habits.

OK, so we've made it accessible, now what? Well, you need to be able to organise it and search it. For organisation, we use tagging and of course hyperlinks, which let you create your own index pages for various purposes.

With a web based model of storage, you open up a lot of options for searching. But, when you are talking about data and facts, rather than text based documents, then you want to be able to run more structured searches and that means having some kind of structure. That brings us on to our next point: understanding the data.

In an ideal world, the person who created, or knows about or 'curates' the data - who we'll call the 'data producer' - gets together with the person who wants to use it, the 'data consumer' and they can talk about it: what the data means, where it came from, what it's limitations are, how it relates to a particular business process or scientific activity. At best,

| | order | | | book | |
|---|---|---|---|---|---|
| | | date order placed | number | total cost (£) | item ordered | price (£) |
| 1 | 0001 | 2008-10-01 | 2 | 9.98 | Green Eggs and Ham | 4.99 |

Fig.1

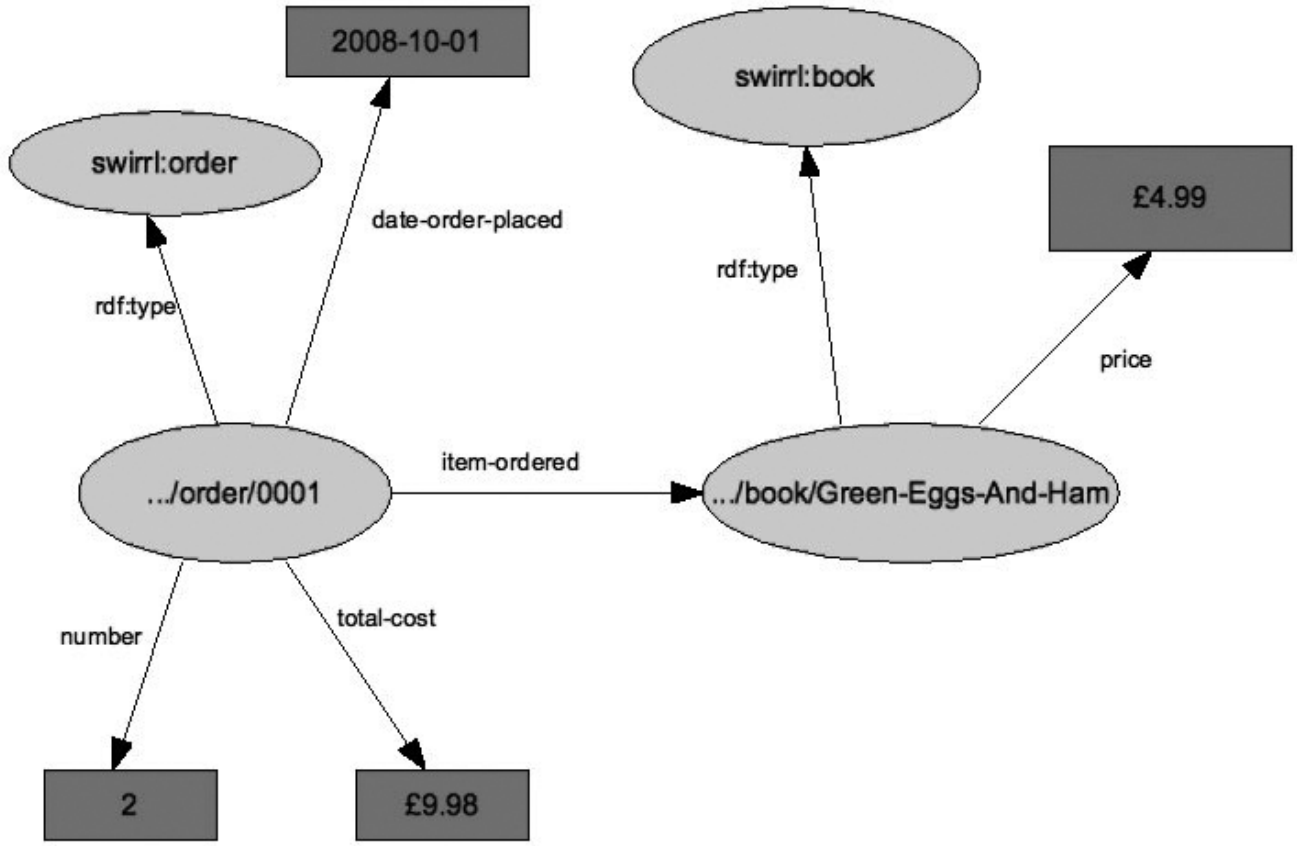



Fig.2

this is a time consuming way of doing things and usually it's not possible at all, because the data producer and data consumer are separated in space or time. Sometimes the data consumer doesn't know who the data producer is: or they are not contactable because they are too busy to respond, have moved jobs, left the company, retired, died or otherwise made themselves difficult to get hold of.

So you need to write down what the data means. What we would really like is to combine data from different sources with the minimum amount of effort: for that to happen those sources have to share some kind of overlapping data model. You need to compare apples with apples, not apples with oranges. If that data model can be stored in some kind of machine processable way, then that reduces the amount of human effort and opens up all kinds of new analysis possibilities.

This is where the semantic web comes in. In our early prototypes of Swirrl, we tried various things: XML + XML Schema, and a kind of JSON-style serialised software objects approach, but we finally settled on RDF as the simplest and most flexible way to store data on any topic. We needed a system of unique identifiers - so that if you and a colleague are talking about the same thing, you call it the same thing. In RDF everything gets a URL, so you have a system for making references to your data at the finest grained level.

We wanted a simple way to represent information: and although the RDF/XML format sometimes does its best to make it seem complicated, representing information as a collection of RDF 'triples' is about as simple as you can get, and fundamentally quite intuitive. You describe something by listing its attributes and the values of those attributes. The graph, in the mathematical sense of nodes joined by edges, is a flexible way of showing how things are related, less restrictive than relational or hierarchical structures. And RDFS and OWL provide a way of describing types and relationships between types, in a standardised language.

*RDFS and OWL provide a way of describing types and relationships between types, in a standardised language.*

The difficult part has been how to make it easy and intuitive for people to enter data in this format. We started from the idea that people like their data in tables: spreadsheets, tables in text documents, or database tables. And when you have lots of data following a similar pattern, a table is an effective way to present it. We wanted as far as possible to 'insulate' users from the fact that we are using RDF behind the scenes, to make it simple to use the system with

the minimum of specialist knowledge. However, if users want to explore and browse through their 'semantic graph' we want to make that possible too.

*We make the simple link that each row of a table represents a 'thing', or 'resource' in RDF-speak and each of the rows in a data set represents a thing of the same type.*

So we organise our data in data sets. A data set is a table of data embedded in a web page. We make the simple link that each row of a table represents a 'thing', or 'resource' in RDF-speak and each of the rows in a data set represents a thing of the same type. Each column represents a property of that resource. Each cell in a table is therefore an RDF statement: the cell holds the value, the row is associated with the subject resource and the column is associated with the property. The value of a cell can be either a literal value, or a reference to another resource.

We also allow properties of linked resources to be included in the row, so a row can in fact be a linearised graph fragment representing multiple resources. (See Figure 1 and Figure 2, illustrating the same data as a row in a Swirrl data set and as an RDF graph). When you give a name to a thing or a property, we use that as a unique identifier in a namespace associated with your wiki. So you can hold information about one thing in lots of different data sets. To try to encourage a group of users to use consistent naming for things and properties, we use autocompletion when the user is typing such names. Nonetheless it is inevitable that some items will end up with different names in different places, and we plan to add more tools for marking things as equivalent to each other, and other ways of 'refactoring' the data.

Our idea is that this should allow the data model behind the data to be as simple or complex as is required and to grow and evolve as people add and work with their data. The power of RDF comes with the ease with which you can extend the data model: if you need to model the entire world up front, then you'll never start.

The end result of this should be a collection of data sets that are interlinked, allowing you to explore the data from different perspectives and to run queries across it. It also provides a sound base for exporting the data to other systems or publishing it: we're currently working on Linked Data compliant publishing of RDF data, with a more comprehensive API to follow.

We think the wiki approach has some big advantages for working with data: when it comes to explaining the meaning and context of your data, there's always a place for good old-fashioned documentation. We tackle this via regular text-based wiki pages, and by making it easy to link between text and data (in both directions). This makes it easy for anyone in the group to add to and improve documentation as they work with the data and to make new links. One feature in the pipeline is to add blog-style comments to data sets, allowing a discussion about the data to take place and be recorded for providing additional future context. These should help make a trail that future users can follow to help them understand and work with the data further.

*We tackle this via regular text-based wiki pages, and by making it easy to link between text and data (in both directions).*

We launched the first version of Swirrl in late September 2008, so we're still in the very early days. We're gathering feedback on which aspects of the system are easy to work with and which not: and to identify which new features are most important for us to add. There's a long way to go, but we think we're heading in an interesting direction.

## Talis Podcasts

**Nodalities**
Mark Greaves talks with Talis about Vulcan and Semantic Web opportunities

In our latest podcast I talk to Mark Greaves, Director of Knowledge Systems Research at Vulcan. We talk about Mark's background, Vulcan's interest in the Semantic Web, and explore the opportunities to build a business on top of semantic technologies.

**Semantic Web Gang**
November/December 2008: The Semantic Web Gang discusses Glue, and looks back at 2008

In the November/December edition of the Semantic Web Gang we discuss the recent launch of Glue from AdaptiveBlue, and look back at the Semantic Web highlights of 2008.

AdaptiveBlue's Alex Iskold is a regular member of the Gang, and shares some of the rationale behind the approach adopted with Glue.

Listen, too, to hear Gang members' perspectives on the events, trends and companies that excelled in 2008… and those with work still to do.

**www.talis.com/podcasts**

# Getting connected

*Justin Leavesley thinks everything's getting connected.*

**By Justin Leavesley**

Web 2.0, social networking, cloud computing, SaaS, PaaS, Web 3.0, the Semantic Web, Smart Phones, 3G, wifi, convergence…. the list of buzzwords or memes goes on - meme being the buzzword for buzzwords.

There is nothing new in a long list of industry buzzwords. However, I think this list is different. It comprises a set of ideas which is each huge and transformative in its own right. The fact that they are all happening more or less at once and are all interconnected should give us serious pause for thought.

Maybe they are better considered as symptoms of some deeper, more fundamental change. It is tempting to focus in on a single symptom and try to understand what that will mean for the future - perhaps even take a risk and build a new business around it. But to focus on a single aspect is to miss the bigger picture. The interaction of several different trends tends to produce serious game-changing disruption. In this climate, it is dangerous to become myopic.

Here is how I would describe the fundamental shift:

"Everything is getting connected."

Obvious? Just to be sure, let me put it another way: EVERYTHING IS GETTING CONNECTED! And I mean everything. I don't mean every blog, every piece of software, every web page, every database—those are just pieces of which software people think everything is made. I mean everything in the world outside the computer screen.

Since the birth of the computer we have begun to build open, generalised infrastructures. The PC is an open and generalised infrastructure for digitising, processing and materialising data. We use the keyboard to digitise text, a mouse to digitise a set of hand gestures, monitors and printers to turn the data back into physical reality; and software organises all of these processes. After all, software is nothing more than a set of instructions which affects data. But the key word here is generalised. We have built machines for thousands of years but they have tended to address specific needs. The PC is a generalised infrastructure for interacting with digital representations. We might use it to manage content such as pictures, music or video. We might use it to write a novel or a business plan. We might use it to organise a supply chain between people and organisations, track financial information, and assess and analyse inventories.

*After all, software is nothing more than a set of instructions which affects data.*

A generalised infrastructure can reduce or eliminate huge costs involved in getting a job done: factoring out some fixed costs and affecting the residual marginal costs of the project. Another way of saying this is that generalised, open infrastructures have huge spill-over effects. If I buy a computer equipped with MS Office in order to organise my personal accounts, my accounts have cost me maybe £1,000. But, of course, I can now word-process a business plan at a marginal rate (i.e. my time). I can also play a game, listen to music and surf the web. That £1,000 actually buys me a generalised piece of infrastructure for a huge range of tasks and functions. I'll leave further discussion of the economics of the spill-over effects created by generalised infrastructures for another time.

Due to the complex nature of these infrastructures, they work much better together when we can agree on some standards. MS Windows and Intel formed a de facto standard which allowed hardware and software to work well together. This partnership has factored out huge complexity by delivering a set of software instructions and processing power to an end user which enables them to manage their data and content. You may argue how much better the world would have been if this standard had been open rather than proprietary, but the point is that the use is generalised and part of the user's infrastructure.

The internet was the birth of an open, generalised infrastructure for connecting computers. Following this, standards have made these networks work much better and the World Wide Web has provided a set of open standards which made the job of connecting human-readable documents much easier. So, the web has provided a generalised infrastructure for connecting documents.

Yet the web isn't limited to connecting human-readable documents. Although it may have come to be thought of as an extension of the PC, it is actually a generalised infrastructure for connecting data: html, mp3, streaming video, xml, rdf—anything in fact. To date, it has mostly been used for html and media content but that is changing rapidly. Connected data is the next logical step and with that we must think of devices and standards.

Take a look at the list of buzzwords again. We are in the process of building a generalised infrastructure for connecting anything to everything. Wifi, 3G and bluetooth allow any electronic device to join the conversation Smart phones ensure the human being is always connected. Thinking about all the digital devices in your life, I expect most of them are currently disconnected. They have to solve all the problems themselves: user interaction achieved through some obscure buttons and a tiny display with odd symbols. They are conceived to be isolated.

But wouldn't it be much better to program your central heating timer with a nice iPhone app that can react to the fact you have left the office? 10 years ago, it would have been impossibly expensive for a heating manufacturer to build a proprietary system allowing customers remotely to programme and adjust their heating. Now, the generalised infrastructure to connect anything is being built, and the huge fixed-cost barrier is being removed. Adding a wifi connection to the central heating controller and exposing the sensor and input data for third party control is already economically doable.

To illustrate this further, imagine how much more valuable it would be for a mass manufacturer to be a bit more connected with their customers. Why, for example, isn't there a big red help button on my washing machine I can press to talk directly to customer support? The washing machine would know its own model number and any error codes it may be displaying and how to contact help. This morning, I was looking at my washing machine and wondering how to control the temperature. You have to select a specific programme, and each has a certain temperature; but it also boasts a separate temperature dial. Does this override the temperature of the programme? Does it add this temperature to the programme, and I'll end up with washing soup? Why can't I literally press a button on the washing machine and immediately ask someone that question and get an answer?

As a products company, that kind of intimate connection with the actual users of my products could be very valuable. For a start, after getting the same questions from many users, they would undoubtedly redesign the temperature control. If the system malfunctioned, the customer service person could give specific advice for an error code, saving time, complaints and dissatisfied customers. This kind of direct relationship has been impossible up to now. With a generalised infrastructure for connecting everything, however, it becomes practicable.

It would cost very little to put a wifi connection on a washing machine and a little Skype-like piece of software which also relays machine status data. I am sure the question is answered in the printed manual, but I dutifully lost that 5 minutes after opening the box. Further, I would not want to go through the hassle of finding the customer services number, finding the model number then reading all that out to the service staff when the machine itself should know all of that. The difference between the effort of simply pressing a button and having to find and relay all that information is functionally massive.

When you think about it, data and devices are everywhere. But they are not connected and they are dumb: they don't know each other; they don't know me; and many can't even recognise themselves.

Almost by definition, there is vastly more localised personal data in the world than generally useful data. Wikipedia is generally useful, and it's helpful to be able to access a postcode-to-location database. But when I think about the data that is really important to me, it is practically all localised to me.

I would dearly like to have a record of my blood pressure and heart rate from the past year. I would share it with my doctor and maybe my pharmacy. But as useful as it would be, there is no way I would spend the effort of taking my blood pressure and writing it down to put in a computer. However, I do take my blood pressure at home with a digital gauge. It is a device that knows something about me, but it isn't connected. My blood pressure, the temperature of my house, the location of my children, miles before a service on my car, the error code on my washing machine, the channels I watch on television, all the local restaurants I have been to in the last year—these are all more valuable to me, my family or my friends than they are generally useful to the world. Just think of the number of people in the world times the personal data relevant to them. I would hazard a guess that it is vastly larger than the generic data in the world.

You can see this effect on Facebook, iff you look at most of what is "published". It is easy to dismiss Facebook as just being full of useless rubbish. If you do, I expect you have fallen into the trap of thinking that just because something is "published" it is meant for you. In fact, most social network content is not a publication, but a conversation that happens to have been digitised. It is intended and meaningful only to a few. It is the same with data. Most data about what I, my family and friends are doing, the things we have, the places we go and the things we need are localised and personal. They are relevant to the few people, organisations and companies I interact with and I want to choose who gets to know what about me.

Up to this point, connecting the devices and data closest to us has been prohibitively expensive, and only a very few people have ever bothered to gather and use this kind of data in a useful way. As the costs of this technology are driven ever-further down, it is becoming increasingly feasible for anyone to have access to these bits and pieces of their lives in a form which can interact and benefit from the infrastructure we build with our devices, houses, cars and companies.

As EVERYTHING gets connected, it is time to get up close and personal.

# Enabling the Linked Data Ecosystem

*Leigh Dodds explores the Linked Data Ecosystem.*

**By Leigh Dodds**

The Linked Data web might usefully be viewed as an incremental evolution beyond Web 2.0. Instead of disconnected silos of data accessible only through disconnected custom APIs, we have datasets that are deeply connected to one another using simple web links, allowing applications to "follow their nose" to find additional relevant data about a specific resource. Custom protocols and data formats are the realm of the early web; the future of the web is in an increased emphasis on standards like HTTP, URIs and RDF that ironically have been in use for many years.

Describing this as a "back to basics" approach wouldn't be far wrong. Many might dispute that RDF is far from simple, but this overlooks the elegance of its core model. Working within the constraints of standard technologies and the web architecture allows for a greater focus on the real drivers behind data publishing: what information do we want to share, and how is it modelled?

*Working within the constraints of standard technologies and the web architecture allows for a greater focus on the real drivers behind data publishing.*

Answering those questions should be relatively easy for any organisation. All businesses have useful datasets that their customers and business partners might usefully access; and they have the domain expertise required to structure that data for online reuse. And, should any organisation want some additional creative input, the Linked Data community has also put together a shopping list [1] to highlight some specific datasets of interest. This list is worth reviewing alongside the Linked Data graph [2], to explore both the current state of the Linked Data web and the directions in which it is potentially going to grow.

Beyond the first questions of what and how to share data, there are other issues that need to be considered. These range from internal issues that organisations face in attempting to justify the sharing of data online, through to larger concerns that may impact the Linked Data ecosystem. For the purposes of this of article, this ecosystem can be divided up into two main categories: data publishers, who publish and share information online; and data consumers, who make use of these rich datasets.

There is obvious overlap between these two categories: many organisations will fall into both camps, as do we all through our personal contributions to the web. However, for this paper I want to focus primarily on business and organisational participants, and attempt to illustrate the different issues that are relevant to these roles.

## Data Publisher's Perspective

The first issue facing any organisation is how to justify both the initial and ongoing effort required to support the publishing of Linked Data. Depending on existing infrastructure this may range from a relatively small effort to a major engineering task—particularly true if content has to be converted from other formats or new workflows introduced. In "A Call to Arms" in the last issue of Nodalities [3], John Sheridan and Jeni Tennison provided some insight into how to address the technology hurdle by using technologies like RDFa.

But can this effort be made sustainable? Can the initial investment and ongoing costs be recouped? And, if a dataset becomes popular and grows to become very heavily used, can the infrastructure supporting the data publishing scale to match?

The general aim with enabling access to data is that it will foster network effects, and drive increasing traffic and usage towards existing products and services. There are success stories aplenty (Amazon, Ebay, Salesforce, etc) that illustrate that there is real and not imagined potential.

But this justification overlooks some important distinctions. Firstly for some organisations, e.g. charities and non-governmental organisations information dissemination is part of their mission and there may not be other chargeable services to which additional traffic may be driven. In this scenario everything must be sustainable from the outset. Secondly, it also overlooks the fact that the data being shared may itself be an asset that can be commoditised. The value of access to raw data, stripped of any bundling application, has never been clearer, or been easier to achieve. New business models are likely to arise around direct access to quality data sources. Simple usage-based models are

1. http://community.linkeddata.org/MediaWiki/index.php?ShoppingList
2. http://richard.cyganiak.de/2007/10/lod/
3. http://www.talis.com/nodalities/pdf/nodalities_issue4.pdf

already prevalent on a number of Web 2.0 services and APIs—the free basic access fosters network effects, while the tiered pricing provides more reliable revenue for the data publisher.

Software as a service and cloud computing models undoubtedly have a role to play in addressing the sustainability and scaling issue, allowing data publishers to build out a publishing infrastructure that will support these operations without significant capital investments. But few of the existing services are really firmly targeted at this particular niche: while computing power and storage are increasingly readily available, support for Linked Data publishing or metered access to resources are not yet common-place.

This is where Talis and the Talis Platform have a distinct offering: by supporting organisations in their initial exploration of Linked Data publishing, with a minimum of initial investment, and a scaleable, standards based infrastructure, it becomes much easier to justify dipping a toe into the "Blue Ocean" (see Nodalities issue 2 [5]).

## Data Consumer's Perspective

Let's turn now to another aspect of the Linked Data ecosystem, and consider the data consumers perspective.

One issue that quickly becomes apparent when integrating an application with a web service or Linked Dataset is the need to move beyond simple "on the fly" data requests, e.g. to compose ("mash-up") and view data sources in the browser, towards polling and harvesting increasingly large chunks of a Linked Dataset.

What drives this requirement? In part it is a natural consequence and benefit of the close linking of resources: links can be mined to find additional relevant metadata that can be used to enrich an application. The way that the data is exposed, e.g. as inter-related resources, is unlikely to always match the needs of the application developer who must harvest the data in order to index, process and analyse it so that it best fits the use cases of her application.

Creating an efficient web-crawling infrastructure is not an easy task, particularly as the growth of the Linked Data web continues and the pool of available data grows. Technologies like SPARQL do go some way towards mitigating these issues, as a query language allows for more flexibility in extracting data. However provision of a stable SPARQL endpoint may be beyond the reach of smaller data publishers, particularly those who are adopting the RDFa approach of instrumenting existing applications with embedded data. SPARQL also doesn't help address the need to analyse datasets, e.g. to mine the graph in order to generate recommendations, analyse social networks, etc.

Just as few applications carry out large scale crawling of the web, instead relying on services from a small number of large search engines, it seems reasonable to assume that the Linked Data web will similarly organise around some "true" semantic web search engines that provide data harvesting and acquisition services to machines rather than human users. Issues of trust will also need to be addressed within this community as the Linked Data web matures and becomes an increasing target for spam and other malicious uses. Inaccuracies and inconsistencies are already showing up.

The Talis Platform aims to address these issues by ultimately providing application developers with ready access to Linked Datasets, avoiding the need for individual users and organisations to repeatedly crawl the web. Value-added services can then be offered across these data sources, allowing features, such as graph analysis (e.g. recommendations), to become commodity services available to all. The intention is not to try and mirror or aggregate the whole Linked Data web, this would be unfeasible, but rather collate those datasets that are of most value and use to the community, as well as shepherding the publishing of new datasets by working closely with data publishers.

As an intermediary, the Talis Platform can also address another issue: that of scaling service infrastructure to meet the requirements of data consumers without requiring data publishers to do likewise. It seems likely that data publishers may ultimately choose to "multi-home" their datasets, e.g. publishing directly onto the Linked Data web and also within environments such as the Talis Platform in order to allow consumers more choice in the method of data access.

## Conclusions

The bootstrapping phase of the Linked Data web is now behind us. As a community, we need to begin considering the next steps, especially as the available data continues to grow. This article has attempted to illustrate a few from a wide range of different issues that we face. While technology development, particularly around key standards like SPARQL, rules and inferencing, and the creation of core vocabularies, will always underpin the growth of the semantic web, increasingly it will be issues such as serviceable infrastructure and sustainable business models that will come to the fore.

At Talis we are thinking carefully about the role we might play in addressing those issues and playing our part in enabling the Linked Data ecosystem to flourish.

4. http://labs.google.com/papers/bigtable.html

5. http://www.talis.com/nodalities/pdf/nodalities_issue2.pdf

# David Provost Talks with Talis

**By David Provost**

**Paul Miller:** Hello and welcome to this Nodalities Podcast with your host, Paul Miller. Today I talk with David Provost about a report he recently published reviewing 17 Se-mantic Web companies.

David, thank you very much for joining me for this podcast today. You are rather topical and in the news at the moment. Before we get into the reason why you are topical and in the news, can you tell us a little bit about yourself and where you came from?

**David Provost:** Sure, I've been in and around the Semantic Web industry since late 2003 when I wrote my thesis titled "Hurdles in the Business Case for the Semantic Web." That's when I was earning a master's degree at MIT. Since then I've worked in the in-dustry in a couple of different capacities, but usually—well, always—on the business side because the last time I wrote any code was when FORTRAN was big and today that might as well be Sanskrit for all its used for. And one thing has led to another, now here I am. I've written a report and talk history.

**Paul:** OK. "Hurdles in the Business Case for the Semantic Web" as your thesis?

**David:** Yeah.

**Paul:** How long ago was this?

**David:** That was published in 2004. And when I wrote that, it was during the course of a mid-career masters program where people who are of my age and makeup, come into Sloan for a year, full-time program. And I wrote this thesis and invited Tim Berners-Lee to be a reader on this thesis. And he was exceptionally helpful and helped me define a variety of concepts and examples that really, I think, added value to what I was trying to say in my thesis.

**Paul:** OK, and what was the point that the thesis came to? What were the hurdles and are they still there?

**David:** Hmm, yeah. Good question. Some of the hurdles identified were simply compre-hensibility, that's what I call it. When you really dig into the technology, it can be ex-tremely complicated and it's very rigourous when it comes to trying to understand the logic that underlies everything that gets done. So, being able to simply understand what's going on is a central issue and remains one, which is why I wrote my most recent report using plain language, because most people are not deep technologists, but yet they still have a desire to learn and use concepts and knowledge that can be applied productively in their own businesses.

So, the language is important and at the time, there were not a lot of success stories, but one of the things that's changed is the number of instances where this technology is being applied and those are some of the things that I've discussed in "On the Cusp".

**Paul:** OK. And yes, I think that is quite an important point about intelligibility, I mean if the Semantic Web and any other technology is going to hit mainstream adoption, you have to talk to more than just people who are already on the inside?

**David:** [laughs] yes.

**Paul:** I think we sometimes forget that, don't we? So, it is worth remembering and being reminded. OK, so let's jump forward to 2008 and this report: "On the Cusp, a Global Re-view of the Semantic Web Industry". OK, what did you find? Who is the Semantic Web industry?

**David:** Yeah, who is? Well, the easy answer is to say it's a group of vendors, some of whom are profiled in the report. But that's certainly not the end of the story by any means. I think one of the most critical issues is, when these vendors make a sale, what happens to the solutions that they are selling, and we are beginning to see some examples where those products are being used. For instance in Twine; Twine uses a variety of technologies to make its experience possible. When Twine first launched, it seemed to me that it was an interest group or a website for a social networking website for people in the Semantic Web community. And that's really not the case, and I don't think it ever was meant to be that way.

*When Twine first launched, it seemed to me that it was an interest group or a website for a social networking website for people in the Semantic Web community.*

I think that the way it's evolving now is much more along the lines of what Nova Spivack intended when he launched Twine, which is that it becomes an interest network for peo-ple with a wide range of interest. For instance, now people have invited me to a cooking Twine so that we can share recipes.

Now, I'm not an avid cook, but it's one example of how this technology can

draw people into areas where they may have an interest, and I certainly receive other invites as well. But it's a way for people to share information and create links across their interests so that they can be aggregated and presented in a consistent format. Yahoo with SearchMonkey is another interesting play and certainly Reuters or Thomson Reuters with Calais.

*Yahoo with SearchMonkey is another interesting play and certainly Reuters or Thomson Reuters with Calais.*

**Paul:** Yeah. Yes, that's interesting. And I think I turned down that invitation to cookery Twine. It is not my thing at all, I quite happily eat it, but making it is just a pain.

**David:** Yeah.

**Paul:** So, I mean you mentioned three there, there are 17 in the report, we may get on to some of the others in a moment, but you mentioned three, Twine, SearchMonkey and OpenCalais. Would any of them have been possible without the Semantic Web? I mean is the Semantic Web a means to an end or is it absolutely essential?

**David:** I would tend to say that it's essential. Here is a common theme that comes up in the industry, which is to say that, you could do all of these things without the Semantic Web, but you could only do them if you were within a small contained environment. As soon as you want to branch out and include the entire world of content, the entire web of content, then you need something that's far more universal. And that's where the stand-ardisation provided by the Semantic Web comes into play and allows you to integrate and interoperate across so many more different kinds of content, and you can begin to introduce the use of applications as well. The Semantic Web is fundamental ingredient to making this happen on a global basis.

**Paul:** But is that not just the Web? The Web is how we reach out across these communi-ties of interest beyond an application, at the moment.

**David:** Yeah.

**Paul:** What more are we getting?

**David:** Well, what you are getting on the Web is a straightforward, highly visual, inter-active experience. It is great for what I call highly targeted interactions. It could be ma-chine to machine. It could be machine to human, where you simply read content that ap-pears on a page. There is no underlying recombination of data information or, at least, that combination does not occur on the fly. It has to be... You have to download it, work with the data, reformat it, and repurpose it to some degree. Then you have an end product that is usable.

What we are talking about, now, with the Semantic Web is the ability to simply arrive at a page and either create those combinations yourself or experience a series of combina-tions that have been readymade for you.

*The Semantic Web is inseparable from the World Wide Web.*

The Semantic Web is inseparable from the World Wide Web. That is because the Seman-tic Web is an extension of the Web. But it makes certain possible that the Web alone just cannot do.

**Paul:** OK. You look to 17 companies in this report. Clearly there is a larger universe of providers than the 17 that you looked at. That aside, do you think the 17 that you got were representative of the whole? Or are they in some ways a skewed subset?

**David:** They are skewed in the sense that they are more prominent. There are a number of companies that were not... Obviously, there are a number of companies that were not covered. Since I am on the business side of things, and my focus is on the business side of things, I wanted to look at companies that had become established to some degree. Or, I wanted to look at companies that had received significant investments and ventures that had received significant investment.

Companies that are established would be along the lines of TopQuadrant, ontoprise, and a variety of other companies that receive significant investment, which would be companies like Twine or Thompson Reuters' investment in Calais.

So, yes. There is a sense of being skewed in that I am not looking at websites that are be-ing launched by one or two people or products that are being developed by one or two people. I am trying to identify the business players in this industry and talk about them.

**Paul:** Right. OK. Presumably, a Website being launched by one or two people, once it attracts money is fair game?

**David:** Yes.

**Paul:** For example, Google.

**David:** Right.

**Paul:** Yeah. You categorise the 17 companies in your report in various ways. Can you talk about some of that categorisation, to give listeners a feel for how this universe of Semantic Web possibilities pans out? There are some extremes in there, aren't there?

**David:** Yes. There certainly are. OK. There are the two broad categories of vendors ver-sus deployers. That's pretty straight forward. The vendors basically make products. I stayed away

from professional services companies. I wanted to focus on people who were building things. So, there are vendors.

Then, on the other side, are the deployers, the people who are using these products to serve a larger purpose, a larger business purpose.

*The Semantic Web is not simply one thing. It is not just RDF. It is not just a bunch of on-tologies. It is those things and a series of other things as well.*

Then, within that, let's just stay with the vendors for a moment. Taking a look at them, it becomes apparent, very quickly, that they actually do different things.

The Semantic Web is not simply one thing. It is not just RDF. It is not just a bunch of on-tologies. It is those things and a series of other things as well.

I wanted to examine the differences within this group of vendors, and talk about the dif-ferent aspects or capabilities that these vendors and their solutions deliver. I think, for a later date, it would be very interesting to simply take a look at the Semantic Web stack and say, "OK, this vendor placed primarily in this place, and that vendor placed primarily in that place."

But that is probably fuel for another report in its own right.

**Paul:** Yes, I think so. And, presumably, a quite technical one, too...

**David:** Yeah [laughs]. It could be beyond me.

**Paul:** [laughs]. Presumably beyond the audience, too, yes? [laughter]

**Paul:** In this survey of 17 companies, all [of them] betting the farm on the Semantic Web to one degree or another... What did you come away thinking? Are they all mad?

**David:** [laughs]

**Paul:** Are they doomed because there is a recession? Are they the cusp of something that is going to transform everything that we do? Or something in between?

**David:** Yeah. That is a good question. Here is how I look at it: the past several years have served as a period where a variety of companies and these are the vendors that are included in the report have gone through that initial formative stage of saying, "Hey, here is a technology. We think it can do something really good and really valuable." Over the course of the past several years, these companies have gone about identifying where their technology—or they have made real progress in identifying—where their technology can be most effectively use.

That period has passed. That is one reason why my basic message is: The Semantic Web has proven itself as a commercially competitive technology.

*The Semantic Web has proven itself as a commercially competitive technology.*

Are they mad? To be in a startup, by definition, you have to be. The thing is that through a lot time and hard work, these folks are beginning to identify their niche. They are be-ginning to identify how their solutions can be used and used effectively.

I think there are a series of examples in that. For instance, Thompson Reuters' sheer rate of uptake... The sheer rate of adoption in terms of how Calais is being used is phenom-enal. Clearly there is some value there. It appears to me that their bet is poised to pay off handsomely.

Other companies are proving that they have a solution that is attractive. Ontoprise is an-other example of that. Mondeca, for instance, with its forays into content management is another example. Each one of the vendors in there, by and large, has a reliable customer base on which it can base predictable revenues, that's what a business is all about, so that is great.

*I do believe that we are at a very interesting point in time where we are going to go from one plateau and climb an arch of growth to a new level.*

So, this is beginning to gel, and since we have gone through that formative stage, I do believe that we are at a very interesting point in time where we are going to go from one plateau and climb an arch of growth to a new level. And where that level is, I don't really know, but I think we are on our way.

**Paul:** I hope so too. So, one of the things that struck me about the 17, they are all actu-ally billing themselves or most of them are billing themselves as Semantic Web com-panies or Semantic Technology companies. And yet, one of the things I have come across quite a lot, I guess most recently at the European Semantic Technology Conference in Vienna couple of weeks back, was this argument, which I agree with, that we shouldn't be selling semantics and semantic technology to a business audience, we should be sell-ing solutions to problems they face.

**David:** [laughs].

**Paul:** To which semantics maybe part of a solution and yet these success stories you highlight, are actually on the whole doing the exact opposite, if you know what I am say-ing, "come and buy the Semantic Web from me."

**David:** Yeah, I think that is very fair, but here is how I look at it. I am very uncomfort-able simply trying to sell the Semantic Web, if I were on the buying side and someone was trying to just trying to say, "you should use this stuff because it is good." If I was a buyer, if I was a customer, I would think to myself, "you got to be kidding." That's why I wanted to write this report the way I did which was to use very plain language in describ-ing what these products actually do. I minimised the mention of technology, the mention of technical terms, the number of acronyms is limited—this in a field that seems to thrive on producing a new acronym. But my focus was to talk about what these products actu-ally do, what value do these products actually bring to you. And that value is not necessa-rily well that value really is not a Semantic Web kind of value, it is a business value.

*My focus was to talk about what these products actu-ally do, what value do these products actually bring to you.*

Let's go back to Twine, With Twine, what Nova has done is used some degree of seman-tics in developing his website so that now he has a viable media property, where people come to his site, they have a focused set of interests and Nova can count on these people staying on the site for about 15 minutes or so. He can identify the demographic informa-tion and that demographic information is attractive to any potential advertiser.

So what he has done on the business level is he has created an interest network that has identifiable demographics, predictable behaviour and he can turn around and sell that. Now he is using semantic technology to achieve a business purpose and I have just de-scribed what he is doing without referring to SPARQL, RDF, XML or anything else that you might ordinarily associate with the Semantic Web.

So I am a very strong believer in talking about what this technology can do and I leave it to a different constituency to talk about what the technology is.

**Paul:** Right. In your reviews of the 17 companies that are in the report, you are quite—how to put this—you are quite glowing, quite upbeat, about the opportunities facing all of them. You are nice about them.

**David:** I am.

**Paul:** Now without asking you to pick one out and be mean about it, are there problems, are there issues, are there things they need to be watching out for?

**David:** Yes. The answer there is yes. Let me just backup for a minute and explain why the general tilt of the report is glowing as you say. I looked at this report as being an introduction to the industry, an introduction "customers please meet these vendors," "vendors, if you haven't already met each other, please meet each other." So in a sense it is like having a party and as your guests arrive, you introduce them to the other people there.

**Paul:** How about "here is David, here is the smelly one." [laughter]

**David:** Yeah right exactly, you are not going to come along and say, "well this guy is a convicted felon and he is currently facing tax evasion charges." [laughs]. That's no way to introduce anybody, so I have tried to simply create an introduction that is easily con-sumed by all players. And in my opinion, you can't really move on to the deeper stuff and talk about specific problems and specific issues without having this as a foundation. So you say, "OK fine, here is the environment, this is what the lay of the land looks like, if you want to dig deeper and do something that is more critical, which I think would be of great value, you can use this as a backdrop, you can use this as a foundation. But before doing anything, I wanted to just define the lay of the land. And as for specific issues, spe-cific issues are simple and very familiar. It is the constant search for this specific problem that your product happens to be best suited for."

To say, "Oh, Semantic Web technology, it is going to be facing great uptake and the rate of adoption is phenomenal." Well, you know what products are represented in that adop-tion and if your product isn't one of them, who cares. So for those vendors and again for all vendors, the issue is "great, if the technology is being adopted, that is fine, but how do you make sure that your technology is adopted," that is ultimate issue when you are run-ning a business.

**Paul:** Yeah. And I might wait till we have turned the recorder off before asking you which company you were thinking of when you talked about "convicted felons wanted for tax evasion?"

**David:** [laughs] fair enough.

**Paul:** Maybe an exercise for the listener to work out what we were talking about.

*I am a very strong believer in talking about what this technology can do and I leave it to a different constituency to talk about what the technology is.*

**David:** They will have to get with me off line for that one.

**Paul:** Yes [laughs]. OK, so you did the report, you produced it over what, two or three months, wasn't it, over the summer. It is being read. People are commenting on it. You picked 17 companies to focus on. Where next? Deeper with the 17? Add more? Or leave it for someone else to do chapter two?

*I think that there is an opening to start posing more critical questions and engaging in more critical work.*

**David:** Good question. It is unclear at the moment. I do have some very clear ideas about other questions that could be asked, and these become a little bit more critical. Just to talk about the report, itself, for a moment, there are a lot of companies that don't appear. A lot of people have pointed that out to me, very generously. My response is, number one, the project had to be a manageable size. Number two, a number of invitations were sent out. Not all companies agreed to participate, for whatever reason. I am certainly in no position to enforce anyone's participation.

Now, with this out of the way, I think that there is an opening to start posing more critical questions and engaging in more critical work as you have described. Also, I think that there are certain avenues to pursue.

For instance, which companies are pursuing what aspect of the Semantic Web stack? Also one that is intriguing to me, but I think would be difficult to pull off is, what is the track record of mergers and acquisitions in this industry?

Why have the mergers and acquisitions followed the pattern that they have? Is there a pattern? Why has that pattern arisen? And how should vendors think about their com-panies and the future of their companies in terms of being acquired, continuing on as standalone businesses, or "other"?

So, yeah, there are definitely a variety of topics that I think would be of interest.

**Paul:** OK. That would be an interesting thing to explore, I think. How does it get paid for? Do we advertise? Do we sell access to the report? Or does your employer think that it is a good thing?

**David:** Yeah. You know, I had a sort of summer vacation when I was doing this. But as for how this gets paid for, I see this kind of report being written on a freely available basis. I think that the follow-on services for more in-depth questions, a more focused examination along the lines of: "Gee, read your report, would like to spend some time with you to get your insight as to where my company fits in with the rest of this industry, the opportunities that you see, the kinds of business models that may be emerging those are more one on one encounters." This would probably follow a traditional analyst or consulting model.

**Paul:** OK. We have your report. It sounds as if we may have iterations to follow that go into more detail and fill in some of these open questions. We have also had, for example the "Semantic Wave" report that Mills Davis wrote, which takes a slightly different tack.

**David:** Yeah.

**Paul:** Where is Gartner? Where is Forrester?

**David:** [laughs]

**Paul:** Where are the other mainstream analysts? Why aren't they looking?

**David:** I don't know. Many years ago, I did work at a company called Gomez. I was an analyst. We were an ecommerce research and consulting company. Even then, I felt that the mainstream research and analysis business had some serious problems in terms of creating and delivering relevant research that people could use and say, "OK. I under-stand exactly what the issue isor at least I understand it better, and now I can actually do something about it." So, where Forester or Gardner or anybody else is... Man, I don't get it. I think there is a problem. I think there is a more fundamental problem that gets in the way of their focus on this kind of issue. I don't know what it is, though.

**Paul:** Maybe you should be offering to write a report for them?

**David:** Maybe I should stay away from them...

**Paul:** [laughs]

**David:** ... Because I don't think I would be able to. [laughter]

*I know that the analysts are interested. I don't know why the firms at the corporate level are not picking up on this.*

**Paul:** Yeah. Hmm. It would be interesting. And, certainly, I know that Gartner have nib-bled around the topic a couple of times, but they have never really got their teeth into it. I am not quite sure why, yet.

**David:** Yeah. That is my impression, too. What I do know is that I have heard that when Semantic Web companies call up a place like Gardner and offer to do a briefing or some-thing like that, there is a strong response among the analysts at these firms to participate in that briefing and learn. I know that the analysts are interested. I don't know why the firms at the corporate level are not picking up on this.

**Paul:** Yeah. I can agree with that from our end, as well. Maybe we need to have a little think and try to find a way to nudge them. This has been interesting. Thank you, David. I look forward to seeing the followup to the current report. Before we wrap up, do you have any closing comment or closing thoughts on what we have been talking about?

**David:** Yes. It is simple. For anybody who is wondering about this technology, it is now safe to stop wondering. It is now time to start experimenting. There are vendors that have proven solutions. These solutions are proven in corporate production environments that are operated by major firms, all around the world. There is a track record that is being built every day, day by day.

If there is any interest or inclination to experiment or pursue this technology, then there are a series of scenarios that I think the report describes. And there are certainly many more that could be discussed.

But there are some scenarios that provide a starting point for people to think and plan about how this technology might be used in their companies to serve some sort of stra-tegic purpose and to serve as a competitive differentiator. The technology is now at the point where it should be considered as a serious component in enabling corporate growth and progress.

**Paul:** Good. Thank you very much, David. That is a good point on which to end. Thank you for your time, and thank you for your report.

**David:** Thank you, Paul.

**Paul:** Thanks.

# Semantic Clouds?

*Semantic Web and the Cloud have much in common*

**By Paul Miller**

Here on Nodalities we tend to focus upon discussion of the Semantic Web, with an emphasis on the Web bit.

In our writing, as in our coding work inside the company, we remain very conscious that the Linked Data mantra we have taken to heart is part of a broader social, economic and technological picture. The technologies of the Semantic Web stack, too, are just components in a broader toolkit that must be used in building real applications and services that will scale to meet genuine business requirements now and in the future.

Cloud Computing is receiving a lot of press at the moment, doubtless aided by Microsoft's entry into the space with Azure. Interestingly, it is also possible to detect a shift in the language of Cloud Computing's enthusiasts that resonates increasingly strongly with some of our own observations. No longer 'just' a cheap, responsive and scalable adjunct to the corporate data centre, Cloud Computing services are finally being perceived as key drivers toward the Web of Data; even champions of the old model such as Salesforce now see enabling third party access to corporate data as a core aspect of their value proposition.

In a post I've tried to explore this in a little more detail. The worlds of Cloud Computing and the Semantic Web are moving closer together, and some fusion of the two appears inevitable. Over the next few editions of Nodalities Magazine, I'd like to use this space to cover the topic of the Cloud, and I'd welcome your thoughts. You can tweet @talis, or comment on the Nodalities Blog (blogs.talis.com/nodalities).

# LIBRIS - Linked Library Data

*Söderback and Malmsten tell the story behind LIBRIS, the Swedish national catalogue.*

**By Anders Söderbäck and Martin Malmsten**

LIBRIS is the Swedish National Union Catalogue, or, in other words, the main gateway for bibliographic data in Sweden. LIBRIS consists of roughly six million bibliographic records, 20 million library holdings records, and two hundred thousand authority records on authors, titles and subject headings ("Svenska ämnesord"). The LIBRIS system is used for cataloguing by about 170 library organisations. The participating libraries come mainly from the academic sector (i.e. university libraries), but it is also possible to find museums, archives and some public libraries. Last but not least LIBRIS is also the home of the Swedish National bibliography. LIBRIS is created co-operatively, but is hosted and maintained by the National Library of Sweden.

Earlier this year, LIBRIS was published as Linked Data on the web, exposing the entire library state with all its records, links and relations. As far as we know, this is the first union catalogue or national library catalogue to be published in its entirety as linked data. Not counting lcsh.info (which is great, but contains no bibliographic information) it is the first effort by a national library to actually be part of the semantic web. We made this effort using a "data first" strategy, focusing on availability rather than a perfect representation of the database. The reason for this strategy was simple: We did not know the perfect structure for a bibliographic web of data. Neither did we want to spend our time just thinking about this. Libraries are, in our opinion, very good at thinking and rethinking bibliographic data. Actual reworking and restructuring of library data unfortunately seems less common. For us, the best way to get to know a technology—whether MARC records or the Semantic Web—is to get our hands dirty and work with it. Progress comes, we firmly believe, by learning from mistakes and learning to adapt to new environments, not by staying safely at home making up the perfect travel plan.

> *Learning also comes from talking and discussing with other people, which is why we wanted to provide something to talk about as quickly as possible.*

Learning also comes from talking and discussing with other people, which is why we wanted to provide something to talk about as quickly as possible. Trying to move beyond the library community, we wanted to use ontologies that were not library specific. For this reason we used Dublin Core, SKOS, FOAF and Bibliontology. Where no existing ontology seemed applicable (holdings, frbr relations), we made up our own. Using the identifiers from our database (MARC field 001, for all you library people out there), we created cool HTTP URIs for bibliographic and authority records. Following the four rules of Linked data specified by Sir Tim Berners-Lee, we tried to provide anyone who looked up those URIs with as much useful information and as many links to other URIs as possible. For the most part, this meant links to other resources within the LIBRIS dataset. Since we already had a good bespoke mapping of our Swedish subject headings to the Library of Congress subject headings, we generated approximately twelve thousand links to lcsh.info. In keeping with our desire to move beyond the library community, we also included a handful of links to wikipedia and dbpedia. In a not too distant future, we plan to automatically extend the number of these links to about thirty thousand.

All of the above is described in more detail in Making a Library Catalogue Part of the Semantic Web (Malmsten 2008). What is not described in this paper, however, is why we decided to depart into semantic web territory. One could of course attribute this departure to inborn curiosity. As librarians and developers we naturally want to explore new landscapes in the ecology of knowledge. But the reason why we started looking at linked data might just as well be more prosaic. In 2007 we made a major revision of our web OPAC. We did this using an open and collaborative methodology, focusing on user centred design as well as communication with just about anyone having an interest in what we were doing. During this process, we discovered that there was a huge interest in getting access to the data contained in the LIBRIS database in a machine readable way. This interest was not limited to the library community, but was expressed by a lot of people working outside the library space with no experience of library specific protocols such as z39.50 and MARC.

In 2008, having our data available in a machine readable-way feels just as natural to us as making our data available in a human readable interface. In building a library catalogue: our purpose and responsibility is to provide access to library resources. Limiting this access to paths that can only be walked by humans seem to be an unnecessary restriction,

which certainly is not in the best interest of our users. The library catalogue might also be considered a resource in itself, and there is no reason why we should not make this resource available for the public just as we do with our books and other resources. Of course, this can be achieved without creating linked data on the semantic web. Why not, for example, use z39.50, SRU/W or OAI-PMH? The answer to this question is that we make our data available in those ways as well. We have also put up a HTTP based web service which can output LIBRIS data as MARC-XML, Dublin Core, Json, RIS or MODS. If we for some strange reason ran out of other things to do, we could probably spend the rest of our lives just building APIs. This situation called for a more rational approach.

*No information is given about other linked resources.*

A trouble with the above mentioned methods of data access is that they only provide information that is present in the original MARC records. No information is given about other linked resources. When looking up a certain resource in our OPAC, a human viewer is provided with a lot of information that is not present to the machine accessing individual records through a search/retrieve protocol. From this perspective, working with rdf and linked data is a lot more rational, since it makes it possible for us to expose the entire state of our library system in a web friendly way. Ed Summers, who is the person at the Library of Congress responsible for creating lcsh.info, describes this very eloquently in a blog post from the beginning of 2008. API:s, Ed writes, "...all differ in their implementation details and require you to digest their API documentation before you can do anything useful. Contrast this with the Web of Data which uses the ubiquitous technologies of URIs and HTTP plus the secret sauce of the RDF triple." As a alternative to making open API:s for just about every aspect of our data, we hope that allowing our data to be crawled by human and machine alike will allow for new ways of discovery, as well as for people using our data in ways not even imagined by us. The Swedish Government has, as one of the objectives given to us as a National Library, stated that the LIBRIS systems shall be used as a broad information resource for the improvement of information management within research and higher education. Putting as few restrictions as possible on how our data can be used is, we feel, probably the best way to achieve this objective.

When it comes to the improvement of information management, the linking between different datasets made possible by linked data shows a lot of promise for cooperation and interoperability between libraries as well as for bridging the gap between libraries and other knowledge organizations. This is something we have only begun to explore. The aforementioned links to lsch.info are useful for making inferences about relations not present in our system of subject headings from relations present in LCSH. The value of these kinds of inferences will increase exponentially the bigger the cloud of linked library data grows. While we at this time only have a few links to dbpedia, we hope that the upcoming addition of a substantial number of dbpedia links will provide interesting possibilities for us as well as for others. Libraries have, over the last few hundred years years, collected huge amounts of what is usually very good data. Locked up in individual databases, this is good for retrieval of resources belonging to individual libraries. This, it needs to be stated, is not a bad thing! Publishing library databases in rdf but without links to other datasets might be good for individual libraries if only for the opportunity to use SPARQL, which is a query language we have fallen in love with and which we imagine might fulfill any librarians desire for good, exhaustive database querying.

Linked library data provides amazing opportunities for cooperation about, for

*The aforementioned links to lsch.info are useful for making inferences about relations not present in our system of subject headings from relations present in LCSH.*

example, authority data. This way of making use of each others intellectual efforts seem to us an effective way of improving the quality of the individual catalogue, while at the same time improving the quality of the web at large. Authority databases are amazing resources, and probably useful for other purposes than just improving search in the individual OPAC. Exposing library data might also be a good way to get feedback on our data from outside the library community. Visibility and open communication provides good ways of quality improvement, which has been proven again and again in science as well as in society. Libraries might be good data providers and a cloud of linked open library data might give rise to interesting perspectives, exciting new applications and better competition between developers, be they commercial system vendors, ad-funded search engines, or in house library development teams. Such competition probably ends up benefiting our users, because in the end libraries are not about catalogues, bibliographic formats, databases or OPACS. Libraries are about openness, about dissemination and access to research and cultural heritage, about science, memory and democracy. For this reason, libraries need to stop worrying and embrace web standards. A web of data without library participation is a bad thing, not only for libraries but also for the web.

# A conference comes of age

*Tom Heath looks back at ISWC, and tells how a conference comes of age.*

**By Tom Heath**

What are the factors that indicate a coming of age? An increased self-awareness perhaps, or an acceptance and understanding of a broad range of views, even if they contradict your own? If these factors do indicate a certain maturity, then I would argue that the International Semantic Web Conference series has come of age.

Last year's event in Busan, Korea felt like a watershed moment, with an increasing focus on practical applications that exploited Semantic Web technologies, in addition to the highly theoretical papers typically seen at events of this sort. This year's conference in Karlsruhe, Germany, and the seventh in the series overall, maintained this momentum. But more so than previous years I detected a subtle change in the mood of the conference. In addition to a tangible sense of excitement that the Semantic Web was getting ready for the mainstream, I detected a certain pluralism within the community, manifested as a greater openness to divergent views and an increase in attention to topics that might have previously been overlooked.

> *the Semantic Web was getting ready for the mainstream.*

This willingness to express and accept divergent views was apparent to me no more so than in the panel titled "An OWL too far?". This discussion saw senior members of the Semantic Web community openly challenge each others views on the proposed second version of OWL, the Web Ontology Language. Perhaps the views held by the likes of Stefan Decker, Frank van Harmelen and Ian Horrocks have always been divergent on this issue, but seeing the differences of opinion aired so openly was a new experience for me. Far from indicating a damaging lack of unity in the field, I read this as a clear sign that the community can engage in open and constructive debate without throwing the toys out of the pram.

Earlier in the week I had sat on a similarly provocatively titled panel in the OWL Experiences and Directions workshop - titled "How might OWL fail?". As a relative outsider I decided to focus on the OWL community's need to improve its marketing and demonstrate its relevance to the wider world, and expected a degree of hostility to this message. Instead I sensed a slight deflation at the criticism that was quickly followed by a desire to engage with the problem and actively address it.

Perhaps the most powerful sign of how far the Semantic Web community has come was in the entries to the annual Semantic Web Challenge. This year the contest had two tracks: the Open Track, which is analogous to the regular challenge in previous years and has a more established set of judging criteria; and the Billion Triples Track, an attempt to stimulate people to generate value from and add value to increasingly large data sets, with the definition of what constitutes "value" being more open-ended.

The quality in both tracks was exceptionally high, but one feature that ran through most of the finalists struck me in particular - the emphasis on the user experience. Previous challenges have always attracted user-oriented applications as well as backend technologies, but this year felt different. Whether the application was supporting personal aggregation of one's distributed information, as in the Open Track winner Paggr; enabling location-oriented browsing of the Semantic Web on a mobile phone, as in DBpedia Mobile, which took second place in the Open Track; or providing structured

browsing over billions of RDF triples, as in SemaPlorer, winner of the Billion Triples Track; the vast majority of entries recognised the need to both add value to the data *and* provide a compelling user experience over this.

For me this indicates not just an awareness but an acceptance on the part of the Semantic Web community that no amount of research and development at the backend will make a difference if clear user benefits are not delivered. If this serves as evidence that the ISWC series has come of age, then I would argue that along with it so has the Semantic Web community at large. It may have taken some time, but I have no doubt that this maturity has been earned.

*Tom Heath is a researcher in the Platform Division at Talis, an active contributor to the Linking Open Data community, and co-chair of the Semantic Web In Use track at next year's International Semantic Web Conference.*

ESWC2009

www.eswc2009.org

# SAVE THE DATES

**European Semantic Web Conference ESWC2009**

**31st May - 4th June, Heraklion, Greece**

Brought to you by: STI · INTERNATIONAL