

INSIDE

August / September 2009

5 Trends and Barriers

6 The Greatest Challenge Facing IT

9 Building a Civic Semantic Web

11 Waiving Rights over Linked Data

13 Might Semantic Technologies Permit Meaningful Brand Relationships?

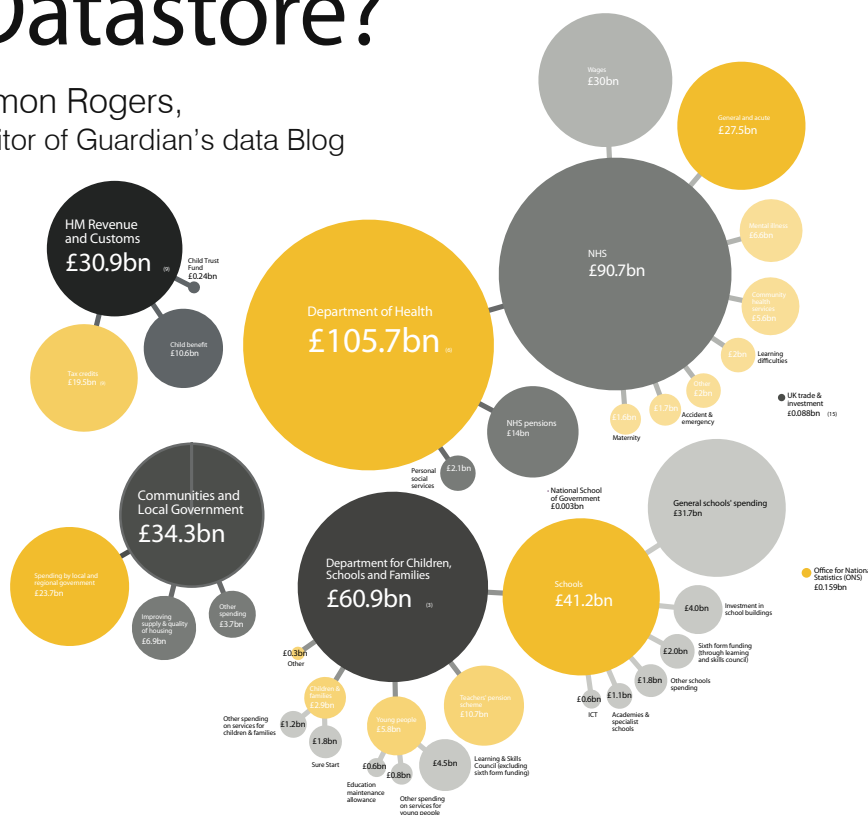
15 RDFa and Linked Data in UK Government Websites

17 Search Engines, User Interfaces for Data, Wolfram Alpha and More...

22 Garlik Announce Open-Source of 4Store

What kind of data is on the Guardian's Datastore?

Simon Rogers,
Editor of Guardian's data Blog



What kind of data is on the Datastore?

- MPs' expenses—full details
- The world in carbon emissions
- The US in plastic surgery
- Deaths in Iraq and Afghanistan
- Breakdown of UK population by race and area
- Public debt since the 1970s
- The world's biggest banks ranked by assets
- The world in refugees
- Interest rates since 1694
- Every Guardian/ICM poll ever

We are drowning in information. The web has given us access to data we would never have found before, from specialist datasets to macroeconomic minutiae. But, look for the simplest fact or statistic and Google will present a million contradictory ones. Where's the best place to start?

That's how our new datajournalism operation came about. Everyday we work with datasets from around the world. We have had to check this data and make sure it's the best we can get, from the

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

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

**FOLLOW US**

Follow Nodalities Magazine on Twitter: @nodalities



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



The views expressed in this magazine are those of the contributors for which Talis Group Limited accepts no responsibility. Readers should take appropriate professional advice before acting on any issue raised.

©2009 Talis Group Limited. Some rights reserved. Talis, the Talis logo, Nodalities and the Nodalities logo where shown are either registered trademarks or trademarks of Talis Group Limited or its licensors in the United Kingdom and/or other countries. Other companies and products mentioned may be the trademarks of their respective owners.

Editor's Notes

Welcome to the seventh edition of Nodalities Magazine.

It's been about a year since I took the mantle of editor, and I've certainly noticed a growth in Semantic Web awareness, building and technology adoption. Perhaps more importantly, however, is a rising awareness of the importance of data to the workings of our societies. From the implementation of data.gov in the US to the promise of

the UK Prime Minister to open up British public data, it's becoming clear that data is growing in coverage and importance.

Following this observation, I'm very happy to include an article which is less Semantic Web, and more data-centric. In fact, it's our cover story this edition, and I think it highlights the importance of publishing data and encouraging its reuse. Simon Rogers, whose background is in putting together info-graphics for the Guardian, writes about his newspaper's Data project which is making information of public interest and record available in data form for people to mashup and reuse.

In order for data to be truly public, however, it is important for an organisation publishing its information to consider the future reuse of its data. Talis' CTO Ian Davis discusses data licensing and waivers—explaining the difference and making it easy to embed machine-readable waivers into published datasets.

Elsewhere, Mark Birbeck explores the potential of RDFa to enable innovation in public-sector websites and government datasets. He gives examples of the UK's Civil Service website, which uses RDFa to embed metadata for job vacancies on the site itself. Alongside this, Joshua Tauberer discusses his ideas for civil/social hacking and the importance the Semantic Web can play in dealing with public-sector information.

Lee Feigenbaum, of Cambridge Semantics introduces Nodalities readers to Anzo, a metadata tool for Microsoft's Excel application. Anzo exists to help with data collaboration in the enterprise, and Lee tells the story behind its development. Paul Miller, from his cloud-computing perspective talks about the Semantic Web and it's relationship to brands in the future. Paul challenges the concept of brand ownership in an era of social media and widely-available data.

Finally, we also have a transcript of a recent podcast in which Paul Miller talks with Garlik's Tom Ilube and Steve Harris about security on the Semantic Web and Garlik's decision to open-source their triple-store (4Store).

As always, I welcome feedback and story submissions. You can follow Nodalities on twitter @nodalities, or me on @zbeauvais. Also feel free to email ideas and feedback to zach.beauvais@talis.com, which also works with MSN Messenger and GTalk.

Zach Beauvais

Continued from front page.



most credible sources. But then it lives for the moment of the paper's publication and afterward disappears into a hard drive, rarely to emerge again before updating a year later.

Comment, as Guardian founding editor CP Scott said, is free. But the second part of his maxim holds equally true for the Guardian today: facts are sacred.

That is where the Data Store and the Datablog come in. Part of the Open Platform—the Guardian's new API—it is about freedom of information.

We like to think it's part of our tradition. Take issue number one of the Manchester Guardian, Saturday 5 May, 1821. News was on the back page, like all papers of the day. And, amid the stories and poetry excerpts, a third of that back page is taken up with, well, facts. A comprehensive table of the costs of schools in the area never before "laid before the public", writes "NH".

NH wanted his data published because otherwise the facts would be left to untrained clergymen to report. His motivation is clear: "Such information as it contains is valuable; because, without knowing the extent to which education, and particularly the education of the labouring classes, prevails, the best opinions which can be formed of the condition and future progress of society must be necessarily incorrect." In other words, if the people don't know what's going on, how can society get any better?

Alternatively sometimes an avalanche of facts is unleashed in order to bury the truth. Journalists have to walk this tightrope everyday, ensuring that the numbers we publish are right.

It's a fine tradition, but one neglected by newspapers until very recently. I hated maths at school. If someone had told me I would spend the most productive part of my professional life working with spreadsheets, I would have been either alarmed or derisory. While journalists accept our mission to inform, we often proudly boasted of our lack of mathematical prowess as if it was a bold character trait. As if the choice at school was either learning maths or English.

And the availability of data reflected that: without computers and accessible numbers, official data was impossible to challenge

accurately. Who knows how many scandals have been obscured by purposely confusing numbers?

But now, publishing data has got much easier. The web has given us easy access to billions of statistics on every matter. And with it are tools to visualise that information, mashing it up with different datasets to tell stories that could never have been told before.

It brings with it confusion and inaccessibility. How do you know where to look, what is credible or up to date? Official documents are

collaborative utopian fantasy."

So, when the commons authorities decided to publish the 500,000 pages of receipts, with very little time for us to process the information, we decided it was time to let some Kool-Aid slurpers loose on it. Software architect Simon Willison came up with an interface that allowed users to rate each of the receipts and add how much they were for. Even better, it was fun to do. The results: over 150,000 pages reviewed in two days, several stories revealed and—crucially—great metadata on how MPs had filed

But now, publishing data has got much easier. The web has given us easy access to billions of statistics on every matter.

often published as uneditable pdf files—useless for analysis except in ways already done by the organisation itself.

Alternatively sometimes an avalanche of facts is unleashed in order to bury the truth. Journalists have to walk this tightrope everyday, ensuring that the numbers we publish are right.

If you're reading this in the US, say, or Canada or Italy or France then Britain's MP expenses scandal may mean very little to you. In the UK, however, the revelations for how British Members of Parliament have been fiddling—it really is the best word—the expenses system have caused nothing less than a shockwave. They have plunged the prime minister's party into shock and resulted in the not-so-swift exit of the Commons speaker, something that has only ever happened once before.

The story was revealed by The Telegraph, which—having purchased the leaked documents early—had 25 journalists working on the project for weeks before any of the stories saw the light of day. It was, said Telegraph commentator Milo Yiannopoulos, a riposte to those who believe anyone can be a journalist:

"Collaborative investigative journalism... feels good because it's messy," said [4ip's Tom] Loosemore, "and could work better than the old models." Oh, yeah? I'd like to see a "messy" collective of Kool-Aid slurping Wikipedians conduct the sort of rigorous analysis necessary for the Telegraph's recent MPs' expenses investigation. Can you imagine social media achieving anything like it? Of course you can't: great journalism takes discipline and training—neither of which exists in Loosemore's

their expenses, claims by party and averages across all categories.

It's not just public data that we're scooping up here. The Guardian produces tonnes of the stuff itself. For instance we did a series of 1000 songs to hear before you die. Based on a spreadsheet and with Spotify links, this is suddenly great data. We also have university ranking tables, raw information about our tax series and so on. In the past, this stuff would have just stayed on the reporter's hard disk. Now, if it's data, it will be on the site.

Increasingly reporters around the world are making it their mission to make data truly free; to publish everything. We have done it with the information behind our series on corporate tax avoidance. Our Free Our Data campaign has even prompted the government to review its information access policies.

So, together with its companion site, the Data Store—a directory of all the stats we post—we have opened up that data for everyone. Whenever we come across something interesting or relevant or useful, we post it up on the site.

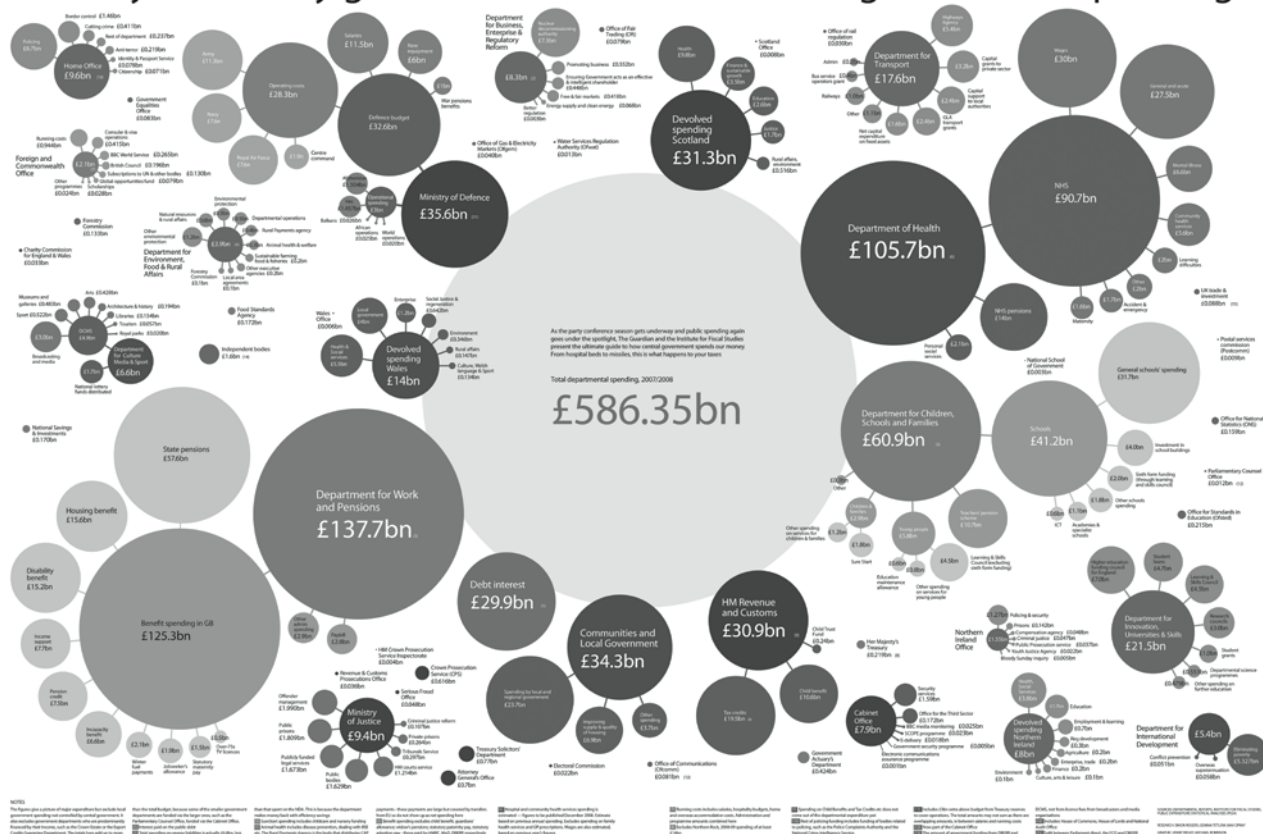
And, rather than uploading spreadsheets onto our server, for now we are using Google Docs. This means we can update the facts easily and quickly, which makes sure you get the latest stats, as we get them. We've chosen Google Spreadsheets to host these datasets as the service offers some nice features for people who want to take the data and use it elsewhere.

It's just a start. We're looking with interest at Google Fusion tables and developing our own visualization apps.



DATA BLOG
Facts are sacred

Where your money goes: the definitive atlas of UK government spending



It is not comprehensive—there will always be things we've missed—but we want our users to help us with that, by posting requests to see if anyone out there knows how to find it. Above all, it is selective. It is information that we find interesting and curious.

A key reason for choosing Google Spreadsheets to publish our data is not just the user-friendly sharing functionality but also the programmatic access it offers directly into the data.

A key reason for choosing Google Spreadsheets to publish our data is not just the user-friendly sharing functionality but also the programmatic access it offers directly into the data. There is an API that will enable developers to build applications using the data too.

We'll be looking at other methods for making data we publish useful both for people and for machines, but we'd love to get some insights from you, as well. Tell us how we can make data more useful.

This is not the end of the road. We need to find ways to make all our data much more useable. Countries are named differently in different datasets, for example. A truly useful data source will have to get round that—probably by using ISO country codes, for example. We're always trying to get round compromises between making our data easy and pleasant to look at and making it useful for developers to take and produce beautiful things with.

But what we have done is to say to the world: we are not some exclusive club that has access to this special stuff that we and only we are going to see.

It is not a one-way process—we want you to tell us what you have done with the data and what we should do with it. As CP Scott said, the facts are sacred: And they belong to all of us.

Simon Rogers is editor of the Guardian's datablog

KEY LINKS:

guardian.co.uk/datablog
A place to get the latest datasets and talk
about those we've put up

guardian.co.uk/data-store
The full directory of all our publicly-available datasets

guardian.co.uk/obamas-america
US datasets

mps-expenses.guardian.co.uk
Our unique crowdsourcing experiment
with the MPs' expenses receipts

Trends and Barriers

By Zach Beauvais



For anyone following the Nodalities blog, you may have read some of my recent posts discussing the trends boiling up around Web 3.0 (other buzzwords are available). The Mobile Web and upgraded connectivity in general; the rise of ubiquitous computing from chips in every product imaginable; Linked Data and the “Semantic Web” as an organising platform for this rising tide of data—these are three very broad trends seeing a lot of media attention presently. From where I’m standing, I tend to see the next great turning point of the Web as a convergence of some of these trends, and see it as a rise in the importance of and reliance upon data itself and data tools generally.

The mobile web is bringing new sorts of information to people, and they can make use of this info wherever they happen to be because of advances in devices and connectivity. As phones and web-enabled devices get better, so do the chips we seem to have embedded all over the place, and we can now begin to have a more clear picture of what we do through the information we gather from our heaters, cars, and pedometers. Also, as more objects become connected, the grunt-work of number-crunching and storage is becoming commoditized into big, efficient, utility-like cloud services, which host and work with our collected information much more effectively than the gadget in your hand could ever hope to do. Others, like us here at Talis, talk about the Semantic Web, which allows for an evolution from a bunch of connected documents to the explicit connections between bits of information.

Also fermenting in this mix is a strengthening trend of political transparency and a public, shared ownership of social data. Barack Obama’s new administration has clearly made this a priority with the launch and work around data.gov; and in the UK, Sir Tim Berners-Lee himself has been appointed to an Parliamentary advisory role. There is growing pressure to be able to have access to public data, and to see it as belonging to the nation’s people rather than allowed to be legitimately filed away in the great, locked bureau of the capitols.

So, picking up two fairly obvious trends here: Social, Public Data and Linked Data; it would seem to follow that people would begin to have access to previously unavailable information in usable, linked forms. And it’s certainly beginning, as articles elsewhere in this magazine have illustrated. But, what about other

chunks of public data? What about when data comes from universities, institutions, scientific foundations and NGO’s? What about charities monitoring crime, CO2 emissions and family histories? Wouldn’t these make a useful piece in the web of social data? What resources have the governments themselves got, if they want to make their public-owned data available in a useful format?

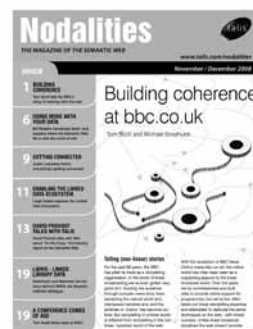
These questions form a major part of the thinking behind Talis’ Connected Commons initiative (talis.com/cc). Basically, Talis has made its Semantic Web platform (including data hosting and access tools) available free of charge for any datasets made available to the public. In doing so, we’re hoping to remove the barrier of cost entirely to publishing interesting data in a Linked Data way. One major reason for this is to promote reuse and mashups of this interesting data, and for people to be able to “follow their noses” to the data that completes their projects. But, from a publishers’ perspective, this is important, because it’s removing a major reason not to bother with making data useful, if not only public. So, with this, data can be made public and useable and the developers and users get the benefit of public SPARQL endpoints and API access to interesting data.

To keep the data open and public, datasets need to make use of either the Public Domain Dedication and License (PDDL) or Creative Commons’ CC0 license. Ian Davis, in his article in this magazine, explains more about waivers and the Connected Commons, and there is a lot more about this particular initiative over on the Talis site (talis.com/platform/cc/faqs/).

In a recent interview with the BBC, Sir Tim said: “This is our data. This is our taxpayers’ money which has created this data, so I would like to be able to see it, please.” I wonder if initiatives such as Connected Commons will begin to remove excuses, hindrances, and obstacles? As public awareness of the importance of access gets hotter, this might become a political issue, as well as a pragmatic one. I hope that in the rush to publish data, and in the ensuing discussion and debate that follows, that the users, hackers and developers don’t get sidelined. I think the world is ready for its data back.

Zach Beauvais is the editor of Nodalities and Platform Evangelist at Talis

Nodalities Magazine 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.

Nodalities Blog

From Semantic Web to
Web of Data

blogs.talis.com/nodalities/

The Greatest Challenge Facing IT

As the old adage goes: time is money.

By Lee Feigenbaum and Mike Cataldo, of Cambridge Semantics



Ultimately, information systems are about saving time. One could argue that technology

enables analysis that facilitates competitive differentiation or improved product quality, but the fact of the matter is that these things and others could all be done without computers; they would just take much, much longer.

A lot has been said and written about information overload. Ultimately, though, the issue with ever-expanding data is that the data we need becomes hidden in mountains of other data. Typically, these mountains take the form of relational databases where the data is neatly stored in rows and columns, and we find the data in one of two ways. Either we directly look up data by its "address" within the database, or else we use a simple text search. But if we don't know what table or column the data resides in, we can't look it up. And as the quantity of data grows, text searching the mountain of data itself yields a mountain of results. Combining through these results then compromises the real benefit of information technology: time savings.

This leads to the greatest challenge facing IT organizations across industries: how to provide users the data they need when they need it, visualized in a way that is understandable and useful. Or put more simply: get the right data, for the right people, at the right time. Traditionally, this is much easier said than done, as the data lives in multiple databases, exists in various formats, and no user interface exists to present the information in a way that is helpful to the user.

Typically, the approach to solving these problems involves some sort of data warehouse. Atop the warehouse, we'd probably deploy a business intelligence (BI) solution to surface the answers to common queries to the people who need them.

Another tactic might be to install a document management system that stores documents in a central repository, where employees can use search and basic metadata to better locate individual pieces of information.

Or we might build a portal to allow people to view the right data from multiple silos in a timely fashion. By defining a collection of portlets as views into specific sources of data, we can provide a one-stop location for people to view information from business-critical data sources.



Pursuing any of these typical solutions means spending 6-18 months at a time solving a single problem. And even worse, all of these approaches are doomed to obsolescence from the start. As requirements change, the fixed schemas and the complex ETL processes inherent to data warehouses must be recreated from scratch. The canned queries and views that define BI- and portal-based approaches must be constantly re-evaluated. And the limited search and query capabilities of a document management system mean that new requirements demand a new installation.

In short, traditional approaches all suffer from the dreaded Shampoo Syndrome: the only workable long-term solution is to constantly lather, rinse, and repeat. And when we do, we just create another mountain of data, another place where what we really need can hide.

The solution is to find data by its meaning rather than its location

The key to eliminating many of the inefficiencies of today's information technology solutions is to access data by its meaning—what it is—rather than its location—where it is. With meaning, we can quickly find what we need simply by describing what it is. This enables information to be shared and consumed at the data level, a paradigm known as data collaboration.

With data collaboration, the data is much more granular, more accessible, and more consumable. In contrast, data warehouse, BI, and portal solutions, in addition to contact tracking (CRM), supply-chain management (SCM), employee management (HR), and all-in-one enterprise bundles (ERP), all fall into the category of data containment. While these applications (commonly known as data silos) excel in capturing extremely structured data, they make it almost impossible to get the data out to be re-used by other users and in other applications.

With data collaboration, the data is much more granular, more accessible, and more consumable.

Document management systems, on the other hand, attempt to make information more shareable, but essentially end up creating many mini-silos in the form of Word documents, PDFs, Excel spreadsheets, or Web pages. This is the world of document collaboration, in which information is readily shared, but the data we need is locked within the min-silo.

Data collaboration is the best of both worlds. By combining the ease of access to information that is the hallmark of document collaboration with the highly structured nature of data from data containment solutions, we can begin to answer the IT challenge. The key to success is to ensure that the meaning of every data element is surfaced so that it can be easily accessed by any person or application that needs it.

Data Collaboration and the Semantic Web

It's no coincidence that the technology standards developed over the past ten years in support of Tim Berners-Lee's vision of a Semantic Web are the key elements for building data collaboration solutions. For as with data collaboration, the Semantic Web relies on explicitly capturing the meaning of data. As such, the core Semantic Web standards pave the way for:

- Flexible, define-as-it-arrives, data structures
- Explicit relationships that travel with the data
- Data that is accessed by its definition rather than its address
- Distributed query

As with all standards, Semantic Web technologies lay the groundwork that makes improvement possible. It is up to application developers to build solutions that make the standards practical.

Practical Data Collaboration to Solve IT's Challenge

Cambridge Semantics is one of the first companies to develop practical business-solution enablers based on Semantic Web

standards. In short, the Anzo products allow businesses to layer a semantic fabric over existing data that:

1. Virtualizes the data so that it is accessible by its description regardless of location.
2. Lets users create their own views of data.
3. Fills in the views by traversing the fabric and picking out the relevant information.
4. Keeps everything in synch by allowing updates that occur anywhere to update information everywhere.

Context. It's not enough simply to have the right data. People must have access to views of the data that depict exactly what they need to see, whether it be an executive dashboard, a regional summary map, or a customer-by-customer detailed report.

The Right Data...

At the heart of the Anzo suite of products is the Anzo Data Collaboration Server. This acts as a central gateway that provides a consistent interface for applications to read, write, and query RDF data, regardless of the actual source of the data. While RDF provides the flexibility to incorporate new data as it is virtualized, it's all for naught without the proper adaptors for existing data sources. To facilitate access to the

right data, the Anzo Data Collaboration Server can connect to data sources including LDAP directories, HTTP-accessible Linked Data, and standard relational databases.

But perhaps one of the most useful connectors is Cambridge Semantics' Anzo for Excel. With Anzo for Excel, data inside spreadsheets with arbitrary layouts can be linked into the Anzo Data Collaboration Server. By breaking down the walls of spreadsheet mini-silos, Anzo for Excel weaves information from thousands (or more) spreadsheets scattered across a business, dramatically increasing the availability of the right data.

...For The Right People

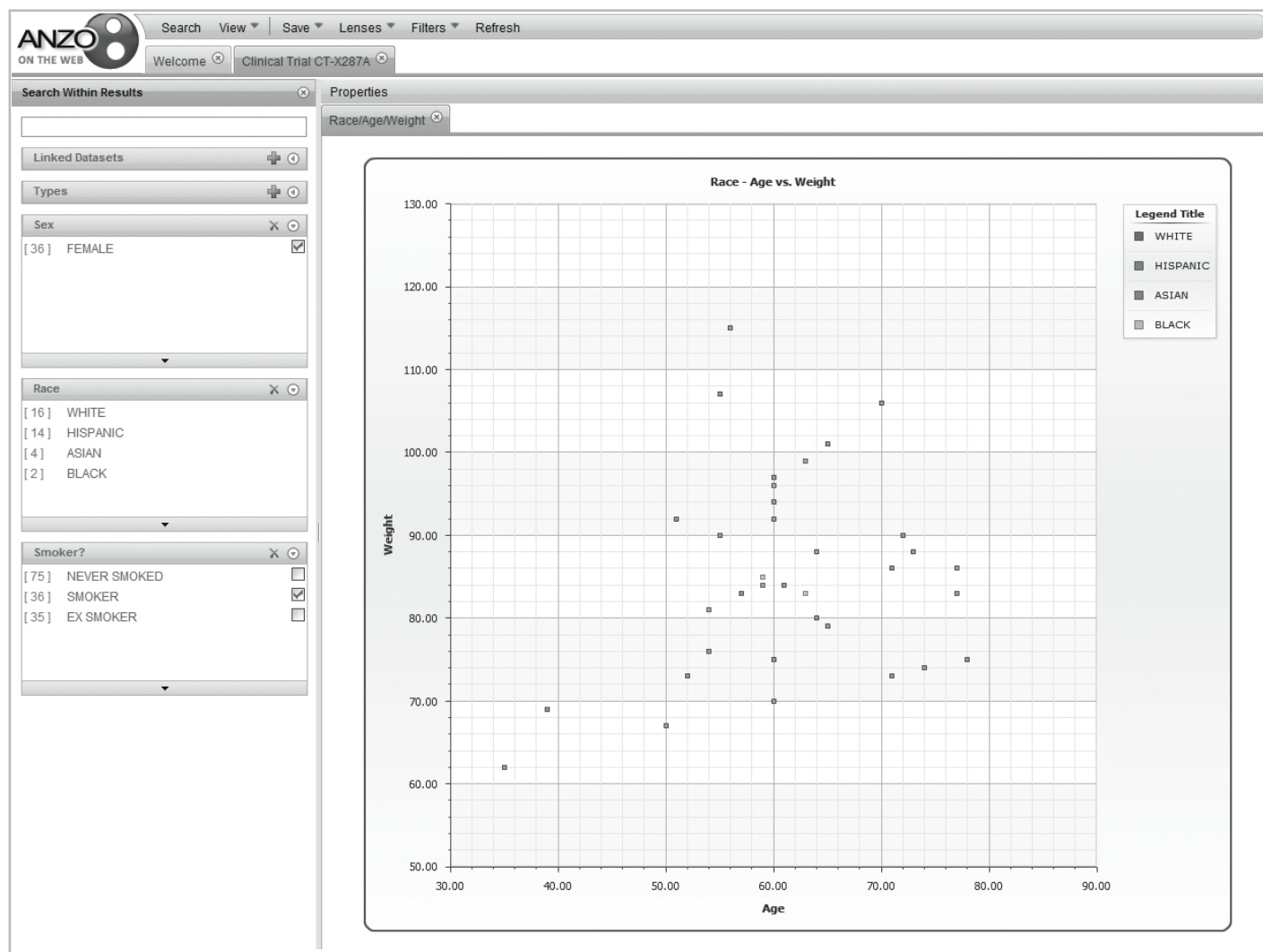
Getting the data in front of the right people relies on three things: context, security, and "reach".

Context. It's not enough simply to have the right data. People must have access to views of the data that depict exactly what they need to see, whether it be an executive dashboard, a regional summary map, or a customer-by-customer detailed report. Cambridge Semantics' visualization product, Anzo on the Web, allows the same information to be rendered in many different ways via semantic lenses. Lenses provide context-appropriate user interfaces to render a particular type of data, meaning that the right people see the right data in the right way.

Security. In many ways, security is the converse of context. While context ensures that the right data surfaces properly to the right people, robust security makes sure data does not surface to the wrong people. The Anzo Data Collaboration Server provides security by layering a role-based access control model atop the semantic fabric. All data access is gated through this security model, which defers to the permissions schemes of legacy data sources where appropriate. The result is that only the right people can ever see (or change) the right data.

Reach. The right data needs to be able to be brought to the right person, whether that person is a technical staff member, a line-of-business manager, a "power user," or a senior executive. As such, the software must be within reach of all users, without the need to call on IT. Research analysts must be able to collect and share spreadsheet data themselves. Anzo for Excel reaches these users by allowing spreadsheets to be visually linked with just a few clicks. Supply-chain managers must be able to drill through data on warehouses, suppliers, and distributors on their own terms. Anzo on the Web reaches these users via a simple and customizable faceted browsing paradigm, whereby anyone can add their own filters, add their own lenses, query their data however they like, and save the results to re-run later or share with colleagues.

| State | FIPS State Code | Jobs Created in Next 2 Years | Total Funding Notification Amount |
|----------------------|-----------------|------------------------------|-----------------------------------|
| ALABAMA | 01000 | 31000 | 1796000000 |
| ALASKA | 02000 | 8000 | 411000000 |
| ARIZONA | 04000 | 70000 | 2274000000 |
| ARKANSAS | 05000 | 31000 | 1100000000 |
| CALIFORNIA | 06000 | 396000 | 13462000000 |
| COLORADO | 08000 | 59000 | 1588000000 |
| CONNECTICUT | 09000 | 41000 | 1436000000 |
| DELAWARE | 10000 | 11000 | 397000000 |
| DISTRICT OF COLUMBIA | 11000 | 12000 | 437000000 |
| FLORIDA | 12000 | 206000 | 6057000000 |
| GEORGIA | 13000 | 106000 | 3420000000 |
| HAWAII | 15000 | 15000 | 511000000 |
| IDAH0 | 16000 | 17000 | 559000000 |
| ILLINOIS | 17000 | 148000 | 5055000000 |
| INDIANA | 18000 | 75000 | 2359000000 |
| IOWA | 19000 | 37000 | 1084000000 |
| KANSAS | 20000 | 33000 | 992000000 |
| KENTUCKY | 21000 | 48000 | 1652000000 |
| LOUISIANA | 22000 | 50000 | 1824000000 |
| MAINE | 23000 | 15000 | 596000000 |
| MARYLAND | 24000 | 66000 | 2046000000 |
| MASSACHUSETTS | 25000 | 79000 | 2924000000 |
| MICHIGAN | 26000 | 109000 | 4066000000 |
| MINNESOTA | 27000 | 66000 | 2029000000 |
| MISSISSIPPI | 28000 | 30000 | 1230000000 |
| MISSOURI | 29000 | 69000 | 2242000000 |
| MONTANA | 30000 | 11000 | 418000000 |
| NEBRASKA | 31000 | 23000 | 759000000 |
| NEVADA | 32000 | 34000 | 868000000 |
| NEW HAMPSHIRE | 33000 | 16000 | 493000000 |
| NEW JERSEY | 34000 | 100000 | 3150000000 |
| NEW MEXICO | 35000 | 27000 | 715000000 |



...At The Right Time

Finally, it's not enough to just bring the right data to the right people. It also needs to be done in a timely fashion.

First, data access against existing data sources is accomplished via federated (distributed) query. SPARQL is explicitly designed to enable queries that access multiple data sources at once, and the Anzo Data Collaboration Server includes a SPARQL engine that does exactly that. By querying the source data directly, Anzo eliminates the cycle time typically associated with a data warehouse's ETL processes.

Second, data updates performed via the Anzo Server are broadcast out in real-time to anywhere the data resides. This means that if a value is changed in a spreadsheet cell, the value instantly updates anywhere else it appears, including Web pages or within a relational database. This is essential as many spreadsheets, Web pages, and databases will share the same piece of data with confidence as semantic tools are made available to users across the business enterprise.

Data Collaboration in the Days to Come

Imagine a world in which this challenge has been solved. End users—whether knowledge workers, line of business managers, or executives—can simply draw a picture of what they want to see and then choose the data that should fill in the picture. Within minutes rather than months the right data shows up on the right people's screens. Now imagine that the data is live as well: you make a correction to the data and your changes are reflected in real-time in whatever legacy database or application the data comes from. You've managed to maintain a single source of truth for your key information assets, while still preserving existing investments in legacy systems and applications.

What sounds miraculous is possible today, in software such as Cambridge Semantics' Anzo. By combining the revolutionary enabling capabilities of Semantic Web standards with solid, practical engineering, we open the door on a completely new paradigm for enterprise software: data collaboration.

Lee Feigenbaum is VP of Technology and Standards and Cambridge Semantics and co-chairs the W3C SPARQL Working Group.

Mike Cataldo is currently CEO of Cambridge Semantics and a veteran of multiple technology start-up companies.

Building a Civic Semantic Web

By Joshua Tauberer of GovTracks.us



Technology is a new key player in government accountability and transparency. It's our own defense against the threat of government information overload. Take the U.S.

Congress: More than 10,000 bills are on the table for discussion at any given time, and Members of Congress are taking campaign contributions from thousands of sources. How can a representative be accountable if his legislative actions are too numerous to track? How can financial disclosure root out conflicts of interest if the interesting ones are buried deep within piles and piles of records? The threat to transparency isn't sheer volume, however. It's the complex network of relationships that makes up the U.S. Congress, and that makes it an interesting case for applying Semantic Web technology.

What the Semantic Web addresses is data isolation, and this is a problem for understanding Congress. For instance, the website MAPLight.org, which looks for correlations between campaign contributions to Members of Congress and how they voted on legislation, is essentially something that is too expensive to do for its own sake. Campaign data from the Federal Election Commission isn't tied to roll call vote data from the House and Senate. It's only because separate projects have, for independent reasons, massaged the existing data and made it more easily meshable that MAPLight is possible. The Semantic Web makes this process cheaper by addressing meshability at the core. The more government data that is mashable, the easier it is to investigate connections across independent data sets, research the dynamics of the system, or teach others how Congress works.

Innovating the public's engagement with Congress by applying technology has been the motivation behind my site www.GovTrack.us, a free congress-tracking tool that I built and have been running since 2004. GovTrack amasses a large XML database of congressional information, including the status of legislation, voting records, and other bits, by screen scraping official government websites that have the data online already but in a less useful form.

If "metadata" is tabular, isolated, and about web resources, the Semantic Web goes far beyond that. It helps us encode non-tabular, non-hierarchical data. It lets us make a web of knowledge about the real world, connecting entities like bills in Congress with Members of Congress, what districts they represent, their

population demographics, etc. We establish relations like sponsorship, represents, voted, and population across entities of many types. A web lets us ask new questions, and from there transforming their answers into visualizations. And because the Semantic Web is a generic platform for all data, I actually think it has the potential to radically and fundamentally transform the way we learn, share information, and live—but that's still a bit far off.

So for the purposes of my tinkering with the Semantic Web, GovTrack creates an RDF dump of its database (13 million triples) covering

`sparql?query=DESCRIBE+%3Chttp://www.rdfabout.com/rdf/usgov/congress/111/bills/h1%3E` using URL rewriting in Apache (for a robust solution, see my explanation at the end of <http://rdfabout.com/demo/census/>). For more about GovTrack's RDF data, see <http://www.govtrack.us/developers/rdf.xpd>.

When data gets big, it's hard to remember the exact relations between the entities represented in the data set, so I start to think of my area of the Semantic Web as several clouds. One cloud is the data I generate from GovTrack. Another cloud is data I separately generate

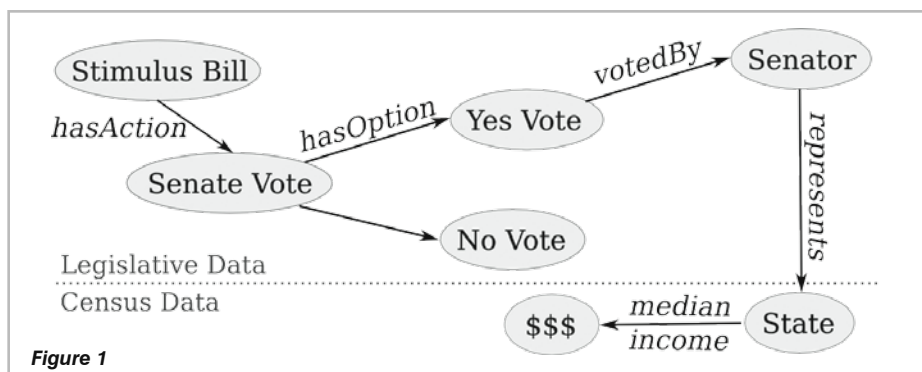


Figure 1

bills, politicians, votes and more using a mix of existing schemas and some new ones that I created. I chose URIs for entities in the Linked Open Data tradition, HTTP-dereferencable URIs that resolve to self-describing RDF/XML about the entity. Two good examples are `<http://www.rdfabout.com/rdf/usgov/congress/people/M000303>` for Senator John McCain and `<http://www.rdfabout.com/rdf/usgov/congress/111/bills/h1>` for H.R. 1, the economic recovery bill passed earlier this year. The HTML pages on GovTrack itself tie in to the RDF world through `<link>` tags: bill pages include the URI I coined for the bill, for instance.

I also have a sometimes-working-sometimes-not SPARQL endpoint set up, SPARQL being the de facto query language for RDF. SPARQL lets us ask questions of the data, such as how did politicians vote on bills (see example 1). The SPARQL endpoint runs off of a "triple store", the equivalent of a relational database for the semantic web, which is underlyingly a MySQL database with a table whose columns are "subject, predicate, object", i.e. a table of triples. (It uses my own C#/.NET RDF library: <http://razor.occams.info/code/semweb>.) The RDF/XML returned by dereferencing the URIs is actually auto-generated by redirecting the user to a SPARQL DESCRIBE query (i.e. <http://www.rdfabout.com/>

about campaign contributions from data files from the government's Federal Election Commission (FEC): 10 million triples. This cloud relates politicians to election campaigns and elections, campaign donors with zipcodes, and contribution amounts. A third data set is based on the 2000 U.S. Census, 1 billion triples. The census data has population demographics for many geographic levels, including states, congressional districts, and postal zipcodes (actually "ZCTA's" but we can put that aside). (For more, see <http://rdfabout.com>. Through the Census cloud the data is linked to Geonames and the rest of the the Linked Open Data community.)

I've related the clouds together so we can take interesting slices through them. The GovTrack data connects to the FEC data through politicians. The Census data connects to the GovTrack data through states and congressional districts (the regions represented by senators and representatives) and to the FEC data through zipcodes. That means we ask questions that go beyond one data set such as: what are the census statistics of the districts represented by congressmen, are votes correlated with campaign contributions aggregated by zipcode, are campaign contributions by zipcode correlated with census statistics for the zipcode, etc.? Once

the Semantic Web framework is in place, the marginal cost of asking a new question is much lower. We don't need to go through heavy work of meshing two data sets for each new question once the data is already in RDF with connected URIs.

My dream is to be able to plug in SPARQL queries into visualization websites like Many Eyes, Swivel, and mapping tools and instantly get an answer to my question in a compelling form. For now, some copy-paste is necessary. Let's take an example. Did a state's median income predict the votes of senators on H.R. 1, the economic recovery bill? Perhaps the senators from the poorest states, likely the most affected by the economic trouble, were more likely to want economic stimulus. This query takes a path through two of my clouds, depicted in Figure 1. The SPARQL query mimics the picture: each edge corresponds to a statement in the query. Except the real query is more complicated (it's given at <http://www.govtrack.us/developers/rdf.xpd>). It is complicated not because RDF or SPARQL are inherently complicated, but because the data model that I chose to represent the information is complicated. That is, I made my data set very detailed and precise, and it takes a precise query to access it properly. If you run it on the SPARQL form on that page, get the results in CSV format, copy them into Excel, and run a correlation test, you'd indeed find a moderate correlation between median income and vote, but in the direction opposite to what we expected. (I know why, but I'll let you think about it.)

Another interesting case is whether campaign contributions to congressmen mostly come from their district, or if they get contributions from sources far away. The SPARQL query listed in example 2 extracts the relevant numbers for Rep. Steve Israel from New York: for each zipcode, the total amount of campaign contributions he received from individuals with addresses in that zipcode in the last election. Figure 2 puts these values on a map, with congressional districts overlaid as well. A form where you can submit a SPARQL query like these examples and see the results instantly on a map would be incredible for data investigation.

So what is government transparency, practically speaking? It's more than just information disclosure. Transparency means the public can get answers to their burning questions. The more questions they can answer from a dataset, the more transparency it provides. We can have more transparency without necessarily more disclosure but instead with the ability to apply better tools. Meshing and querying government datasets with RDF and SPARQL could be a new way to reach new heights of civic engagement and public oversight.

Example 1

Get a table of how senators voted on all of the Senate bills in 2009-2010:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bill: <http://www.rdfabout.com/rdf/schema/usbill/>
PREFIX vote: <http://www.rdfabout.com/rdf/schema/vote/>
SELECT ?bill ?voter ?option WHERE {
  ?bill a bill:SenateBill .
  ?bill bill:congress "111";
  bill:hadAction [
    a bill:VoteAction ;
    bill:vote [
      vote:hasOption [
        vote:votedBy ?voter ;
        rdfs:label ?option ;
      ]
    ]
  ]
}
```

Example 2

Get total campaign contributions to Rep. Steve Israel by zipcode.

```
PREFIX fec: <http://www.rdfabout.com/rdf/schema/usfec/>
SELECT ?zipcode ?value WHERE {
  ?campaign fec:candidate <http://www.rdf...ongress/people/I000057> .
  ?campaign fec:cycle 2008 .
  ?zipcode
    fec:zipAggregatedContribution [
      fec:toCampaign ?campaign;
      fec:amount ?value
    ] .
  ?zipcode fec:zcta ?uri .
}
```

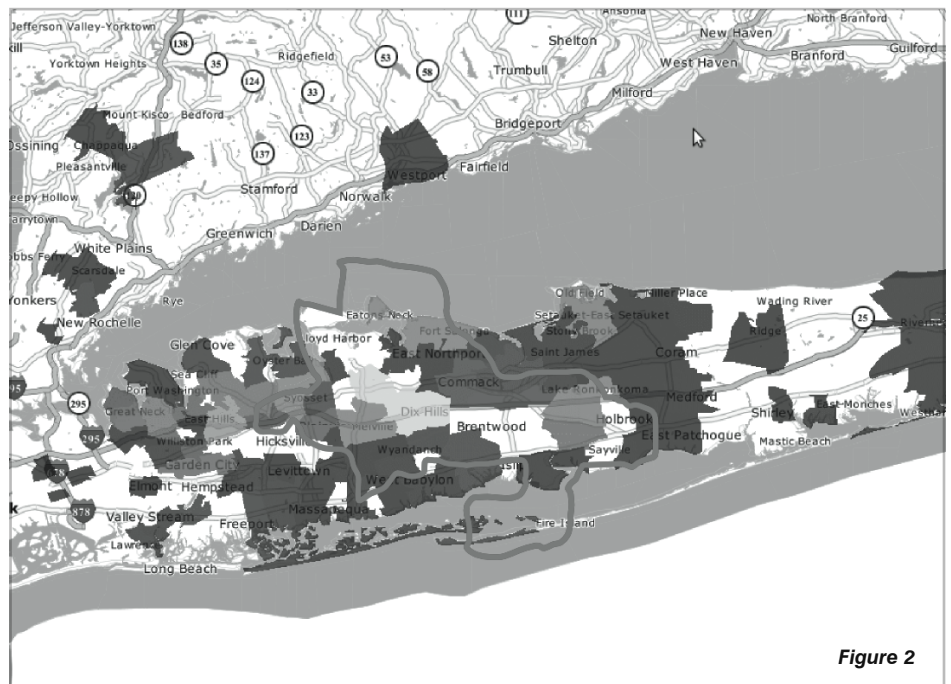


Figure 2

Joshua Tauberer is a software developer and entrepreneur and runs www.GovTrack.us, which tracks what is happening in the U.S. Congress.

Waiving Rights over Linked Data

By Ian Davis, Talis CTO



We love data at Talis and we want as much of it to be freely reusable as possible. In fact, because we wanted to see even more reusable data we recently launched the Talis Connected Commons offering completely free hosting of public domain data. We believe that dedicating data to the public domain is the best way to ensure that data is universally reusable and remixable. When data is public domain it means that it can be reused automatically without needing to check terms and conditions or track the source of every statement to provide attribution. These kinds of things act as friction to reuse, wasting energy that could be better spent creating inspiring things.

We also firmly believe that, in the future, there will be a significant role for other forms of data licensing, including commercial access. We will support those efforts too when the time comes but today the Linked Data web needs more and better data that is freely accessible.

Licensing vs Waivers

If you are not interested in the background to data licensing then you can jump straight to the section How to Declare Your Waiver.

You are probably familiar with the process of licensing a creative work, most likely through the great job that Creative Commons have been doing in recent years. However, the concept of waivers is less well known but highly relevant to reuse of linked data.

Whenever you create something you have automatic rights over it granted to you.

Whenever you create something you have automatic rights over it granted to you. The best known of these rights is copyright, which gives you the exclusive right to make copies of your creative work. There are many other rights which can be held over intellectual property such as design rights, trade marks, registered designs, performers rights, trade secrets, database rights, publication rights and many more.

Licensing is the process of granting others limited use of rights you possess. For example, when you license your copyright you are

granting specific people a limited right to make copies without having to ask you first. Licensing of one right does not affect your possession of the others. For example you could grant the right to copy your work but retain the right to perform it. Creative Commons licenses are mostly concerned with copyright, but they do not usually deal with the other rights such as database rights or trade secrets.

Waivers, on the other hand, are a voluntary relinquishment of a right. If you waive your exclusive copyright over a work then you are explicitly allowing other people to copy it and you will have no claim over their use of it in that way. It gives users of your work huge freedom and confidence that they will not be pursued for license fees in the future.

The Licensing Problem

In general factual data does not convey any copyrights, but it may be subject to other rights such as trade mark or, in many jurisdictions, database right. Because factual data is not usually subject to copyright, the standard Creative Commons licenses are not applicable: you can't grant the exclusive right to copy the facts if that right isn't yours to give. It also means you cannot add conditions such as share-alike.

There isn't a Creative Commons license for every possible right and there probably can't be because of the huge variation in rights granted in different jurisdictions around the world. Also, when we start to look at licensing compilations of data we find that the situation becomes complex because you have to consider both the database and its contents separately. For example a document of articles would be subject to database right over the whole collection and individual copyrights for each article, quite possible to many different owners. The Open Data Commons has addressed this particular example with its Open Database License and Database Contents License (based on work originally donated by Talis). If a standard license doesn't exist then you need to hire lawyers and write one for yourself—a potentially huge cost.

Our collective goal for a successful Linked Data web has to be to protect consumers of the data: the people who are remixing many different sources of data. Our intentions may be very honourable, but people need certainty if they are to build enduring value on data. Creative Commons licenses are irrevocable so even if you lose control over your work through some misfortune, the people reusing it will be protected forever. Imagine this scenario: you

allow people to use data you have collated but your company goes bankrupt and the rights to the data collection are sold by the liquidators. If you hadn't licensed your rights explicitly then every one of your users could be liable to be sued by the new rights holder!

This is where waivers of rights can help. By explicitly waiving your rights over your data then you are giving your users the best guarantee of safety that you can. Even if you lost control of the data collection subsequent owners could not pursue your users because the rights you held have already been waived.

There are two waivers of rights that can be applied to datasets:

- PDDL from Open Data Commons
- CC0 from Creative Commons

Both of these waivers can be used for data intended for submission to the Talis Connected Commons.

Community Norms

When you apply a waiver like CC0 you are relinquishing all your rights over the work to the fullest extent possible under the law. That means that you cannot force people to attribute you or stop them from making commercial use of your work.

The preferred approach is to attach a set of community norms to the work. These are like a code of conduct for use of the work and are usually self-policing. They are not legally enforceable but form part of the ethical or professional requirements for participating in a community. The best known example of community norms are the citation standards used in the academic community. Citing pre-existing work is not legally enforceable but those who abuse the norms can find themselves excluded from the academic community.

The Open Data Commons has published a set of attribution and share-alike norms which asks that users of the data:

- Share work derived from the data.
- Give credit to the original data publisher.
- Point others at the source of the data.
- Publish in open formats.
- Avoid using digital rights management.

How to Declare Your Waiver

To declare your waiver in a machine readable way, you should first create a vOID description of your dataset. vOID, or Vocabulary of Interlinked Datasets, is a vocabulary designed to describe key attributes of your dataset. We created a waiver RDF vocabulary that can be used with

void to declare any waiver of rights and the community norms around a dataset.

In this example we describe a dataset using the void:Dataset class and provide it with a dc:title as a minimal human readable description. You should add other descriptive properties as necessary (some suggestions can be found in the void guide).

We then use the wv:waiver property (defined in the waiver RDF vocabulary) to link the dataset to the Open Data Commons PDDL waiver. We use the wv:declaration property to include a human-readable declaration of the waiver. This is purely informational, but can be immediately be used by a person examining the void description. Finally we use the wv:norms property to link the dataset to the community norms we suggest for it, in this case the ODC Attribution and Share-alike norms.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/terms/"
  xmlns:wv="http://vocab.org/waiver/terms/"
  xmlns:void="http://rdfs.org/ns/void#">
  <void:Dataset rdf:about="{uri of your dataset}">
    <dc:title>{{name of dataset}}</dc:title>
    <wv:waiver rdf:resource="http://www.opendatacommons.org/odc-public-domain-dedication-and-licence"/>
    <wv:norms rdf:resource="http://www.opendatacommons.org/norms/odc-by-sa"/>
    <wv:declaration>
      To the extent possible under law, {{your name or organisation}} has
      waived all
      copyright and related or neighboring rights to {{name of dataset}}
    </wv:declaration>
  </void:Dataset>
</rdf:RDF>
```

Alternatively if you were to choose the CC0 waiver without any particular norms then you should use the following RDF:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/terms/"
  xmlns:wv="http://vocab.org/waiver/terms/"
  xmlns:void="http://rdfs.org/ns/void#">
  <void:Dataset rdf:about="{uri of your dataset}">
    <dc:title>{{name of dataset}}</dc:title>
    <wv:waiver rdf:resource="http://creativecommons.org/publicdomain/zero/1.0"/>
    <wv:declaration>
      To the extent possible under law, {{your name or organisation}} has
      waived all
      copyright and related or neighboring rights to {{name of dataset}}
    </wv:declaration>
  </void:Dataset>
</rdf:RDF>
```

These examples show that it is very simple to declare your waiver. However, before you do so be sure to read carefully what rights you are irrevocably giving up. For example you would most likely be waiving your publicity and privacy rights, so if your image is included in the dataset you could not later complain that someone is using it in a way you do not approve of. If you are worried about how your work will be used, if you want to legally require attribution, or if you don't want people to make money off of your work, then you should not use a waiver and instead seek legal advice on the creation of a data license specific to your needs.

Might Semantic Technologies Permit Meaningful Brand relationships?

By Paul Miller, Founder of cloudofdata.com



Much has been written about growing Enterprise use of social media (usually Twitter, these days) to successfully track and mitigate customer complaint. Many have been quick to spot that the disproportionately high cost of satisfying (or, more cynically, silencing) these early adopters is unlikely to scale effectively as an increasingly large cohort of customers move onto these services, and it must remain an open question as to whether ComcastCares and its peers can survive any move to the mainstream in recognisable form.

Collectively, we've moved from simply complaining about the worst failures of companies, their products and their employees, toward emitting an impressive stream of FYIs.

It appears, though, that Enterprise engagement in the social sphere changes the game far more significantly than merely enabling a select few twitterati to jump the Customer Support queue, and that this change is worth effort and investment in order to ensure that it does scale. What's actually happening is that a relationship is being enabled between a brand and those that Seth Godin might recognise as its tribe; a relationship in which interactions are no longer driven predominantly by the desire to seek redress. Rather than only raising those issues serious enough for us to have written letters or endured telephone muzak in the past, we now comment on issues at the periphery of a brand. Collectively, we've moved from simply complaining about the worst failures of companies, their products and their employees, toward emitting an impressive stream of FYIs. Individually insignificant, and possibly unimportant, together these light touches on and around a brand build into an ever-changing and valuable commentary that brands and the corporations they front would do well to take notice of. The minor niggles

about an otherwise exemplary service, the human touches that made us smile, the odd inconsistencies in a polished persona; none are enough to make us pick up the phone, but we comment upon them endlessly in Twitter, Facebook, FriendFeed and elsewhere, and by tapping into this fundamentally honest stream of consciousness there is much for those about whom we comment to learn. Good companies probably already know about fundamental failings in a product long before their customer support operation melts down under the weight of complaints or their quarterly sales targets are seriously under-achieved. Do they have as good a handle on the things we love? Do they have a clue about the minor gripes of customers outside their pre-launch polling groups? Do they know about the gut reaction to a colour, a touch, a smell, or a careless word that persuaded a likely prospect to buy a technically or aesthetically inferior product from the competition instead? All this and more is there for the taking in the stream of online chatter freely directed their way.

Semantic Technologies aren't often directly associated with the worlds of marketing and commerce, yet individuals such as Eric Hillerbrand and Scott Brinker are hard at work to show just what might be possible when the experiences of the Semantic Web are applied to this space. Brands are no longer owned by the companies in whose name they were created. Increasingly, ownership of various forms is being asserted by the multitude of stakeholders with effort and attention invested in the brand. They care about it, they care about what it says about them, and they play a clear role in the brand's evolution whether its managers want them to or not.

Brands are no longer owned by the companies in whose name they were created. Increasingly, ownership of various forms is being asserted by the multitude of stakeholders with effort and attention invested in the brand.

Brands need to engage in this conversation, as we are beginning to see them do, but they also need to discover the means to cost-effectively monitor and engage with a potential flood of third party reaction whilst using the Business Intelligence tools available to them in nimbly shaping public opinion to their advantage wherever possible.

I spoke with Scott Brinker last year (<http://bit.ly/wlkwp>), to explore his—then nascent—views on Semantic Marketing, and look forward to hearing his latest thoughts at the Semantic Technology Conference in San Jose in June.

More recently, Eric Hillerbrand (<http://bit.ly/XyAA9>) talked about some of his ideas with respect to 'Social Commerce,' and the ways in which commercial organisations might seek to strengthen and exploit relationships with their customers, aided by a range of semantic technologies.

We're just beginning to grasp the realities of a world in which tightly controlled and fiercely guarded brand attributes become increasingly permeable.

We're just beginning to grasp the realities of a world in which tightly controlled and fiercely guarded brand attributes become increasingly permeable. For those companies with the confidence and foresight to loosen their grip, whilst simultaneously exploiting the wealth of data and new opportunities to engage, there is much to be gained. For the dinosaurs that hang on to 'their' brand in spite of the world around them, there is everything to lose.

Paul Miller is the founder of The Cloud of Data blog.

Liberate your data and let a thousand ideas bloom.



Talis provides a Software as a Service (SaaS) platform for creating applications that use Semantic Web technologies. By reducing the complexity of storing, indexing, searching and augmenting linked data, our platform lets your developers spend less time on infrastructure and more time building extraordinary applications. Find out how at talis.com/platform.

shared innovation™



RDFa and Linked Data in UK Government Websites

By Mark Birbeck, of Backplane Ltd



The UK government's Central Office of Information had a straightforward problem to solve: how could they create a centralised website of information that the public could search and access, when the source of that information could be any government department database or any public sector website?

For example, different organisations, such as Her Majesty's Revenue and Customs (HMRC) or the National Health Service (NHS) would each post job vacancies to their own websites, but

these consultations are on departmental sites, rather than being available on a central site; from the Department of Energy and Climate Change (DECC) seeking feedback on clean coal, to the Ministry of Justice (MOJ) providing an opportunity for people to comment on prisoners' voting rights, each department manages its own publication of consultations.

Traditional solutions

Traditional answers to these problems would have been to either (a) impose on each of the departments that they should key their data directly into a new central database (which

development work, retraining of users, porting data from older systems, and so on.

The second 'traditional' solution at least has the merit of keeping existing systems intact, but would have required additional interfaces to be created to move the data from the departmental servers to the centre; each department would have had to create an interface between their own system and the central one.

Just getting one department into a situation where they could centralise their information would have been a major undertaking—not only were there lots of departments to consider, but each department was using a different technology to publish their vacancies or consultations to the web. For example, some departments with only a small number of job vacancies would likely use static HTML pages. Other departments, perhaps with larger IT departments, might use ASP.NET or a Java-based system.

Enter RDFa

The RDFa answer to this set of problems is simple—both conceptually, and to implement.

RDFa allows HTML publishers to embed RDF into their pages, so using the HTTP and HTML infrastructure to publish their information. This simple method of publishing data in turn means that any system can import this data, just by obtaining (or creating) an RDFa parser.

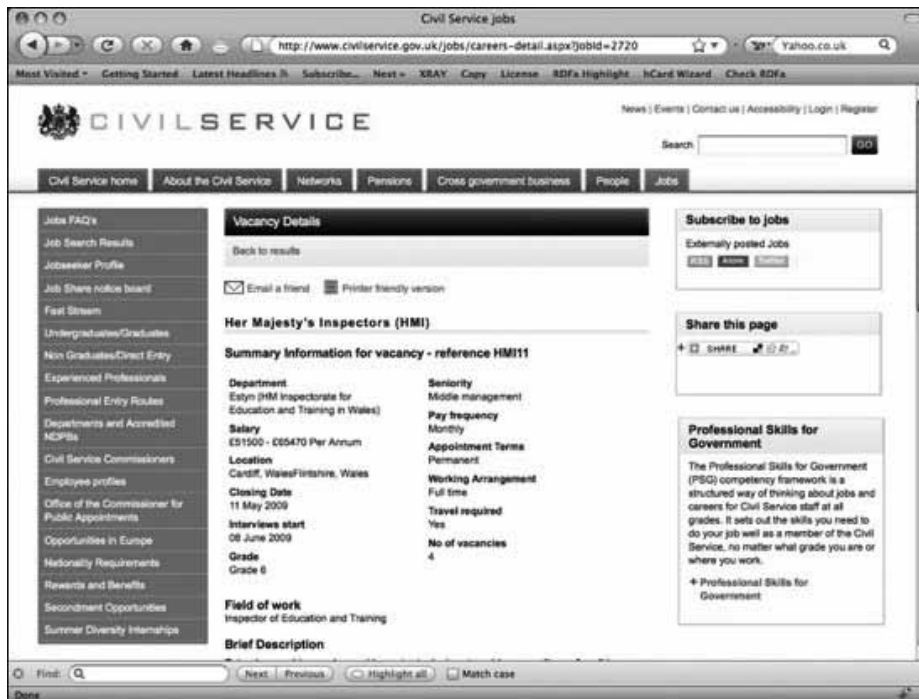
In short, each department can keep their own data management system, and simply add code to their existing web-page publishing step to augment the HTML with the data as RDFa. The central system in turn only needs one import mechanism—something that understands RDFa.

Adding this facility to an individual departments publishing system proved to be very quick and straightforward. But it's not just UK government departments that are finding it straightforward to add RDFa to their pages. It was interesting to hear at SemTech in June that Google's rich snippet launch partners (such as Yelp), were able to add RDFa support in "roughly a day".

RDF publishing techniques

Adding data to web-pages might seem quite an obvious technique, but there are two important things to note here.

First, the COI has to be commended for having the vision to publish RDF at all. Of course, now that Gordon Brown has asked



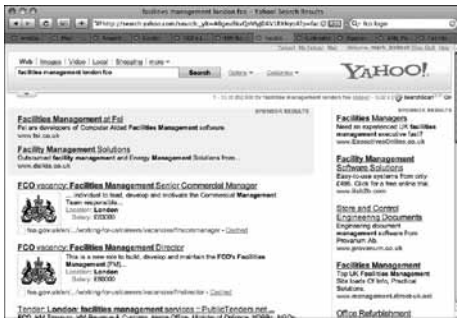
there was no central site that the public could go to, to find all public sector vacancies. This would be a problem at any time, but in the midst of attempts by the government to help people through the recession, it's crucial to ensure that the public knows what vacancies are available. It might not occur to someone looking for a job as a plumber or an electrician they should visit the NHS or Army websites, so a centralised site could make a big difference.

Similarly, as in most modern democracies, government departments are constantly seeking feedback from the public and interested parties, about specific issues. But as with job vacancies,

would in turn drive the central website), or (b) create complex communication pipelines that would allow the decentralised databases to communicate with the central system.

And either of these solutions would almost certainly have turned out to have been a non-starter.

The first solution was unlikely to ever get off the ground, because it would have required each department to replace their existing technology with something new. Even if there was agreement on what that technology should be—and that in itself could take an age to resolve—there would have been a need for new



for Sir Tim Berners-Lee's help in making government data publicly available, it seems pretty obvious—indeed it may even become fashionable! But the COI were planning this project at least a year ago, and at that time RDF was by no means a done deal (and you could say it's still not).

But the second important thing is that even after deciding to publish RDF, it's still not immediately obvious that the solution should involve RDFa, especially not a year ago.

The usual means of publishing RDF is to provide a distinct source of data in the form of RDF/XML (and perhaps other formats, too, such as N3). If there is an HTML version it usually exists for the purpose of describing the data itself. In other words, the RDF/XML format is primary, which means that anyone who is publishing HTML pages but wants to publish RDF as well, will need to add an extra piece of infrastructure that exists alongside their web-pages.

RDFa turns this on its head, and says that the HTML page is the data. One and the same page can be read as an HTML page, or as an RDF page, which in turn means that the changes required to the existing publication system are minimal. The COI once again showed its far-sightedness by adopting this technique.

Turtles all the way down

But the benefits of RDFa don't just stop there. Firstly, because the data is being published via HTTP and HTML, it's possible for anyone to read the same data, not just the centralised website that was being planned. This means that third party job vacancy sites, for example, could import vacancies from relevant departments, to add to their databases. In fact, one of the main drivers for the consultations project was to try to help improve the accuracy of an already existing website (set up by a member of the public) that used 'screen-scraping' to try to keep up with the available consultations—RDFa provides much more accurate information.

In addition, the centralised website will not only import RDFa but publish it too. This means that third-party servers are also able to import some or all of the centralised data, into their own sites.

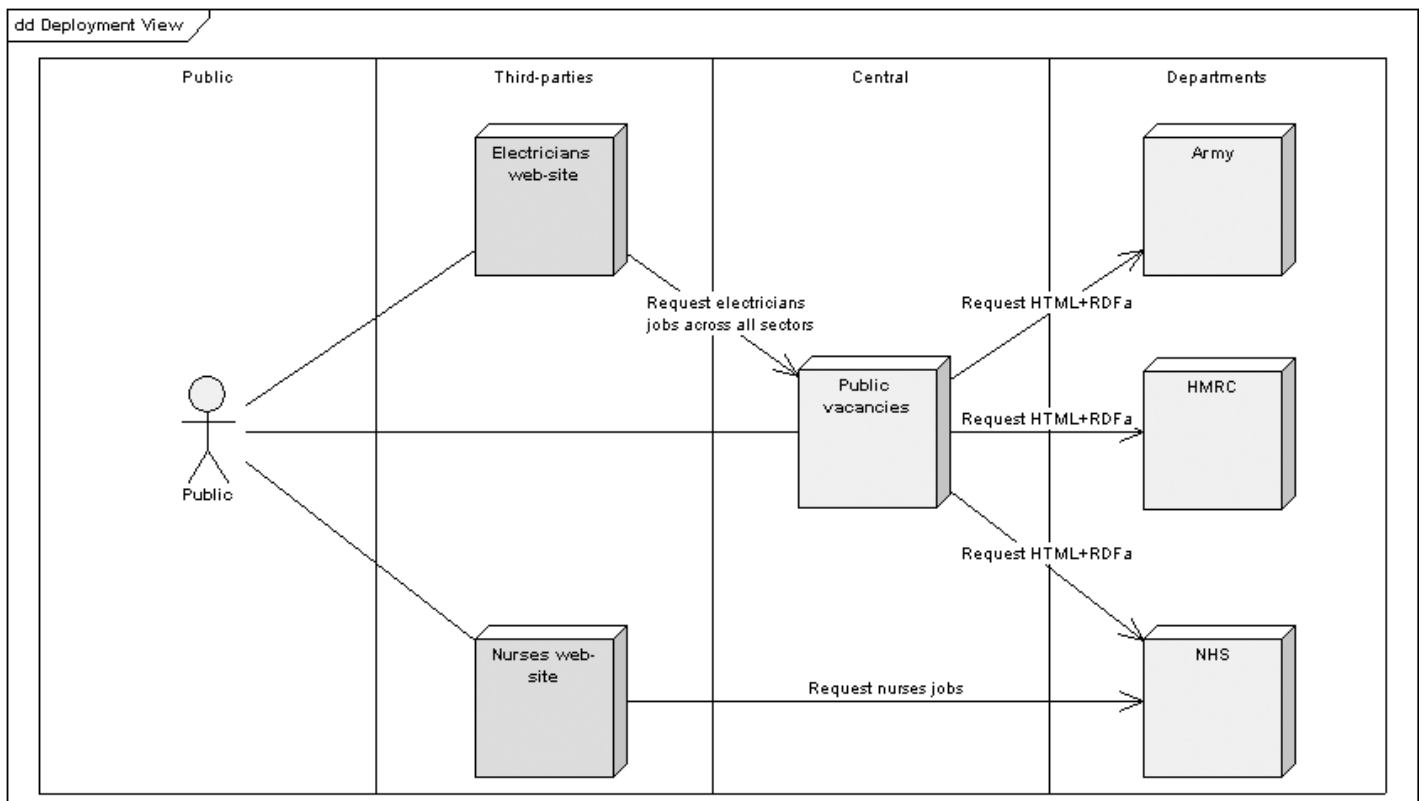
And thirdly, by using RDFa the sites could provide information to search applications such as SearchMonkey.

As more servers both consume RDFa from one set of servers, and publish RDFa again to a variety of other servers, we enter the exciting world of Linked Data, and it's 'turtles all the way down'.

Conclusion

By using RDFa to address the challenge of making distributed data available in one place, the COI avoided having to make changes to each department's systems. But once each department is publishing RDFa, it becomes possible for third parties to consume that information however they see fit. Such a flexible architecture is crucial in the age of open government, and is a cornerstone of linked open data.

Mark is managing director of Backplane Ltd. (<http://webBackplane.com/>), a London-based company involved in a number of RDFa/linked data projects for UK government departments. He is the original proposer of RDFa.



Search Engines, User Interfaces for Data, Wolfram Alpha, and More...

Interview of Sir Tim Berners-Lee by Richard MacManus of ReadWriteWeb. Reprinted with kind permission.



During my recent trip to Boston, I had the opportunity to visit MIT. At the end of a long day of meetings with various MIT tech masterminds, I made my way to the funny shaped building where the World Wide

Web Consortium (W3C) and its director Tim Berners-Lee work. Berners-Lee is of course the man who invented the World Wide Web 20 years ago.

This was my first meeting with the Web's creator, whose work and philosophy was

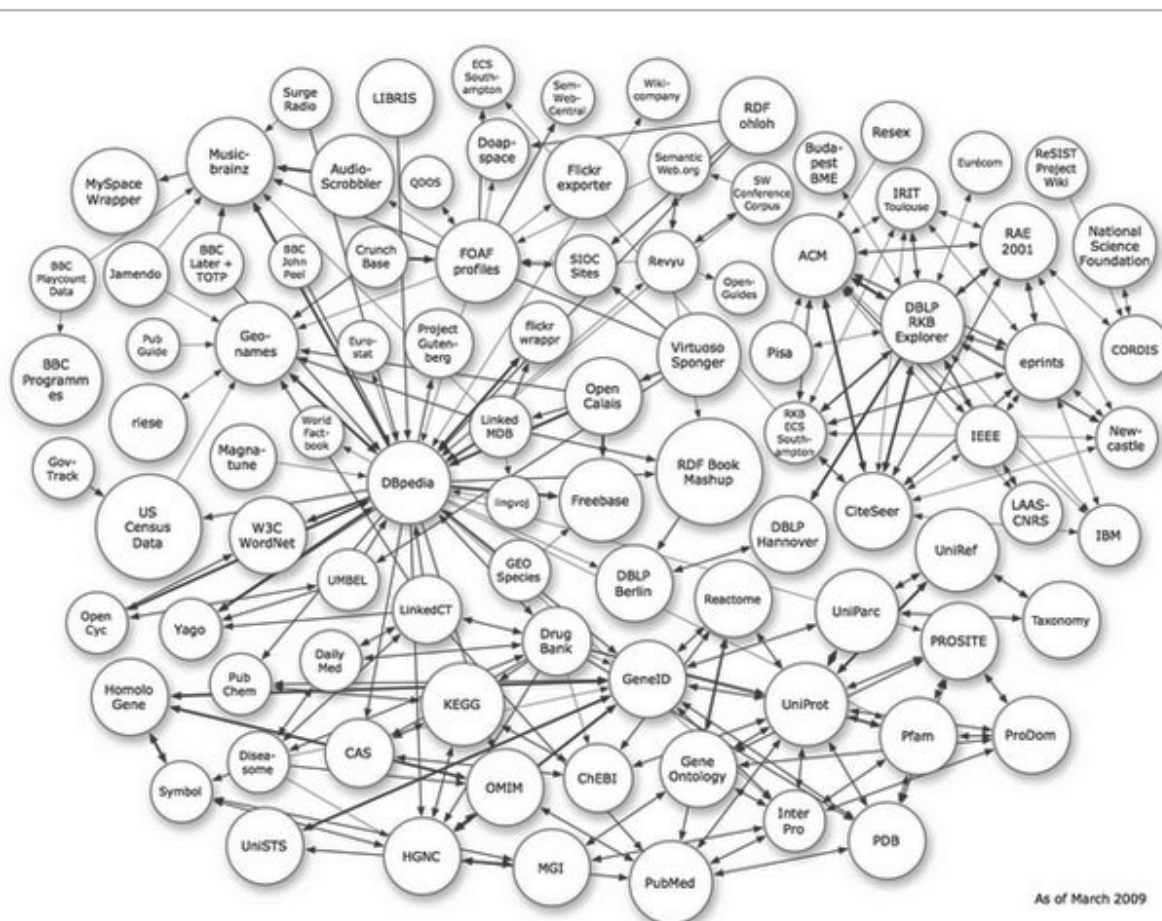
a direct inspiration for me when I launched ReadWriteWeb back in 2003.¹

After shaking hands, I told Tim Berners-Lee that this blog's name was in part inspired by the first browser, which he developed, called "WorldWideWeb". That was a read/write browser; meaning you could not only browse and read content, but create and edit content too. It was a shame then when Mosaic, a read-only browser, became the first mainstream web browser in the mid-90s. It wasn't until the rise of Web 2.0 that the read/write philosophy gained widespread acceptance.² On that note, we launched into the interview...

How Linked Data Relates to The Semantic Web

RWW: Earlier this year you gave an inspiring talk at TED about Linked Data. You described Linked Data as a sea change akin to the invention of the WWW itself - i.e. we've gone from a Web of documents to a Web of data. Can you please explain though how Linked Data relates to the Semantic Web, is it a subset of it?

TBL: They fit in completely, in that the linked data actually uses a small slice of all the various technologies that people have put together and standardized for the Semantic Web.



As of March 2009

1. The very first sentence written on this blog, on 20 April, 2003, was: "The World Wide Web in 2003 is beginning to fulfill the hopes that Tim Berners-Lee had for it over 10 years ago when he created it."
2. For more on read/write browsers, you can read another early RWW post entitled 'What became of the Browser/Editor'.

We started off with the Semantic Web roadmap, which had lots of languages that we wanted to create. [However] the community as a whole got a bit distracted from the idea that actually the most important piece is the interoperability of the data. The fact that things are identified with URIs is the key thing.

The Semantic Web and Linked Data connect because when we've got this web of linked data, there are already lots of technologies which exist to do fancy things with it. But it's time now to concentrate on getting the web of linked data out there.

How Linked Data Has Evolved via Grassroots

RWW: Linked Data has had a lot of grassroots support, which you mentioned in your TED speech. This is something Semantic Web technologies, such as RDF, have struggled to get over the years. Has the W3C been pushing the more bottom-up Linked Data world, because of the frustration over lack of take-up of top-down Semantic Web?

TBL: A lot of the initial RDF and OWL projects came out of the academic world; and some of them were projects to show what you could do in a closed world. And the files were zipped up and left on a disc. While they were interesting projects, and while the systems were useful systems, the semantic web community maybe missed the point of the 'web' bit and focused too much on the 'semantic'. However the work that's been done in the Semantic Web, the standards, was really valuable. It's relatively recently for example that SPARQL [an RDF query language] has been developed.

Somebody drew an analogy the other day: can you imagine trying to promote a world of databases without SQL? Even though it's not an interoperable protocol, it's just a query language. So similarly, all that's been put into RDF, rdfs and OWL is very valuable to the linked data community.

The Linked Data community tend to use a subset of that [Semantic Web technologies], of OWL for example. But they certainly use SPARQL. So you could argue that really it wasn't ready to be deployed widely.

Linked Data started as a very informal Design Issues note that I put in; it was a grassroots movement from very early on. So yes W3C has been emphasizing the importance of Linked Data. It's been the Semantic Web Interest Group of course, and various [other Semantic Web] activities, which has been pushing it. But also Linked Data has been seized on - a group of people for example put together DBpedia.³ That wasn't commissioned, that was that they just thought it would be a really cool idea.

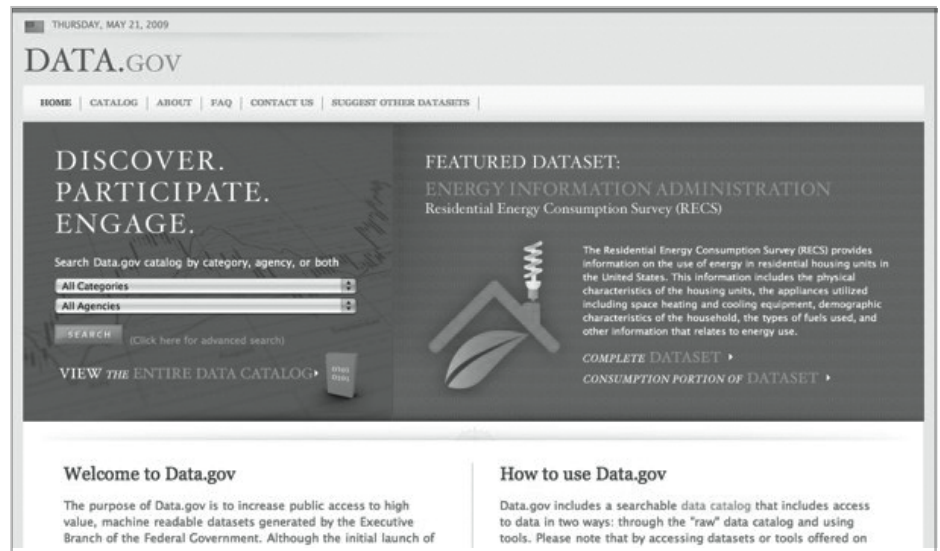
Linked Data and Governments

RWW: In a recent Design Issues note, you urge governments to put their data online as Linked Data (although you'd also be happy for governments to just make available the raw data - presumably so that others can then structure it). What do you realistically expect, for example, the U.S. or U.K. governments to do over the next year? And in the near future, do you foresee different governments interconnecting their Linked Data sets?

TBL: One can't generalize, governments

[their data] into Linked Data. One of the nice things about Linked Data, when they have a pile of it, is that they could run a SPARQL server on it. SPARQL servers are a commodity product, a solution for all of the people who say 'but actually I wanted to have XML.' A SPARQL server will generate an XML file [and] allow somebody to write out, effectively, a URL for the XML file.

In fact, I don't see why SPARQL servers shouldn't provide CSV files, something which as far as I know isn't in the standards. But I'd



are (like most big organizations) fascinatingly diverse inside them. So you'll find that there are places inside governments where you get a champion who gets linked data and who's just written a script and produced some linked data. So in the UK government for example, you'll find there's RDFa [in the code of its website] for civil service jobs. So if somebody wants to make a database of all the jobs, they can do that very easily.

There are other cases where the easiest thing for somebody to do is to just put data up in whatever form it's available. Comma separated values (CSV) files are remarkably popular. They're exported sometimes from spreadsheets. It's remarkable how much information is in spreadsheets. Or sometimes pulled out of a database and then put up on the web. It's not as good, not as useful to the community, as if Linked Data had been put up there and linked. But the first step of actually putting the data out there is the one that nobody else can do.

Data.gov, a catalog of public data, was launched in May by the U.S. government

The way to go is for government departments to go the extra step and convert

recommend it, certainly in government context, because CSV files are what people have and what people want.

So the message [for government] is to use RDF. Linked Data is the backplane, it's the thing that you connect to in both directions. As a [web] producer your job is to make sure that you produce Linked Data one way or another. And as a consumer, there are lots of ways to consume that data once it's out there as Linked Data.

Semantic Web and Search Engines Like Google, Yahoo

RWW: You've been talking about the Semantic Web for many years now. Generally the view is that Semantic Web is great in theory, but we're still not seeing a large number of commercial web apps that use RDF (we've seen a number of scientific or academic ones). However we have begun to see some traction with RDFa (embedding RDF metadata into XHTML Web content), for example Google's Rich Snippets and Yahoo's SearchMonkey. Has the takeup of RDFa taken you by surprise?

TBL: Not really, but the takeup by the search engines is interesting. In a way I was happy to

3. DBpedia is a community project to extract structured information from Wikipedia; see ReadWriteWeb's profile of this and similar resources.

see that, it was a milestone for those things to come out of the search engines. The search engines had typically not been keen on the Semantic Web - maybe you could argue that their business is making order out of chaos, and they're actually happy with the chaos. And if you provide them with the order, they don't immediately see the use of it.

different thing to be able to do. It really doesn't compare to a search engine.

You've got search for text phrases on one side (which is a useful tool) and querying of the data on the other. I think that those things will connect together a lot.

So I think people will search using a search text engine, and find a webpage. On the front

interfaces or designs for semantic applications in recent months - if so which ones and why did they impress you?

TBL: I think that whole area is very exciting at the moment. The only piece of hacking I've done over the past few years has been on a thing called the Tabulator [a data browser and editor], which is addressing exactly that. Partly because I wanted to be able to look at this data. And now there are lots of different ways that people need to be able to look at data. You need to be able to browse through it piece by piece, exploring the world of data. You need to be able to look for patterns of particular things that have happened. Because this is data, we need to be able to use all of the power that traditionally we've used for data. When I've pulled in my chosen data set, using a query, I want to be able to do [things like] maps, graphs, analysis, and statistical stuff.

So when you talk about user interfaces for this, it's really very very broad. Yes I think it's important. There's also the distinction we can make between the generic interfaces and the specific interfaces.

There will always be specific interfaces; for example if you're looking at calendar data, there's nothing else like a calendar that understands weeks, months and years. If you're looking at a genome, it's good to have a genetics-specific user interface.

However you also need to be able to connect that data, through generic interfaces. So if my genome data was taken during an

Also I think there was misunderstanding in the search engine industry that the Semantic Web meant metadata, and metadata meant keywords, and keywords don't work because people lie. Because traditionally in information retrieval systems, keywords haven't proven up to the task of finding stuff on the Web. One of the reasons is that people lie, the other is that they can't be bothered to enter keywords. So keywords have gotten a bad reputation, then metadata in general was tarred with this 'keywords don't work' brush. Because a lot of Semantic Web data included metadata, then people thought that with Semantic Web data -- again, that people will lie and won't have the time to produce it.

Now I think there's a realization that when you're putting data online, that people are motivated NOT to lie. For example when your band is going to produce its next album, or when your band is going to play next downtown, you're motivated to put that information up there on the Semantic Web. There's an awful lot of cases when actually data is really important to people; and it's on the web anyway. So I think it's great that some of the search engine companies are starting to read RDFa.

Does this mean that they [search engines] will start to absorb the whole RDF data model? If they do, then they will be able to start pulling all of the linked data cloud in.

Will they know what to do with it? Because when it's data in a very organized form, I think some people have been misunderstanding the Semantic Web as being something that tries to make a better search engine - i.e. when you type something into a little box. But of course the great thing about the Semantic Web is that you can query it, you can ask a complicated query of the Semantic Web, like a SQL query (we call it a SPARQL query), and that's such a

of the webpage they'll find a link to some data, then they'll browse with a data browser, then they'll find a pattern which is really interesting, then they'll make their data system go and find all the things which are like that pattern (which is actually doing a query, but they'll not realize it), then they'll be in data mode with tables and doing statistical analysis, and in that statistical analysis they'll find an interesting object which has a home page, and they'll click on that, and go to a homepage and be back on the Web again.

URI: <http://www3.wiwiiss.fu-berlin.de:2020/all>
This is release 0.7 of the Tabulator Project. The [live development trunk](#) is also available.

- The Tabulator Project
- Decentralised Information Group
- W3C (large: includes all specs and groups)
- W3C groups (Member-only data)
- Life Sciences demo data: GSK3-Beta Drug Target
- ▼ D2R Server contents
 - label D2R Server contents
 - seeAlso
 - ▼ List of all instances: Conferences
 - mentions ► Conference
 - label List of all instances: Conferences
 - seeAlso ► D2R Server contents
 - ▼ International Semantic Web Conference 2002
 - location Sardinia
 - date June 9-12, 2002
 - type ► Conference
 - label International Semantic Web Conference 2002
 - seeAlso ► List of all instances: Conferences
 - startDate 2002-10-09T00:00:00.0
 - is conference ► Trusting Information Sources One Citizen at a Time
 - of ► Automatic Generation of Java/SQL based Inference Engines from RDF Schema and RuleML

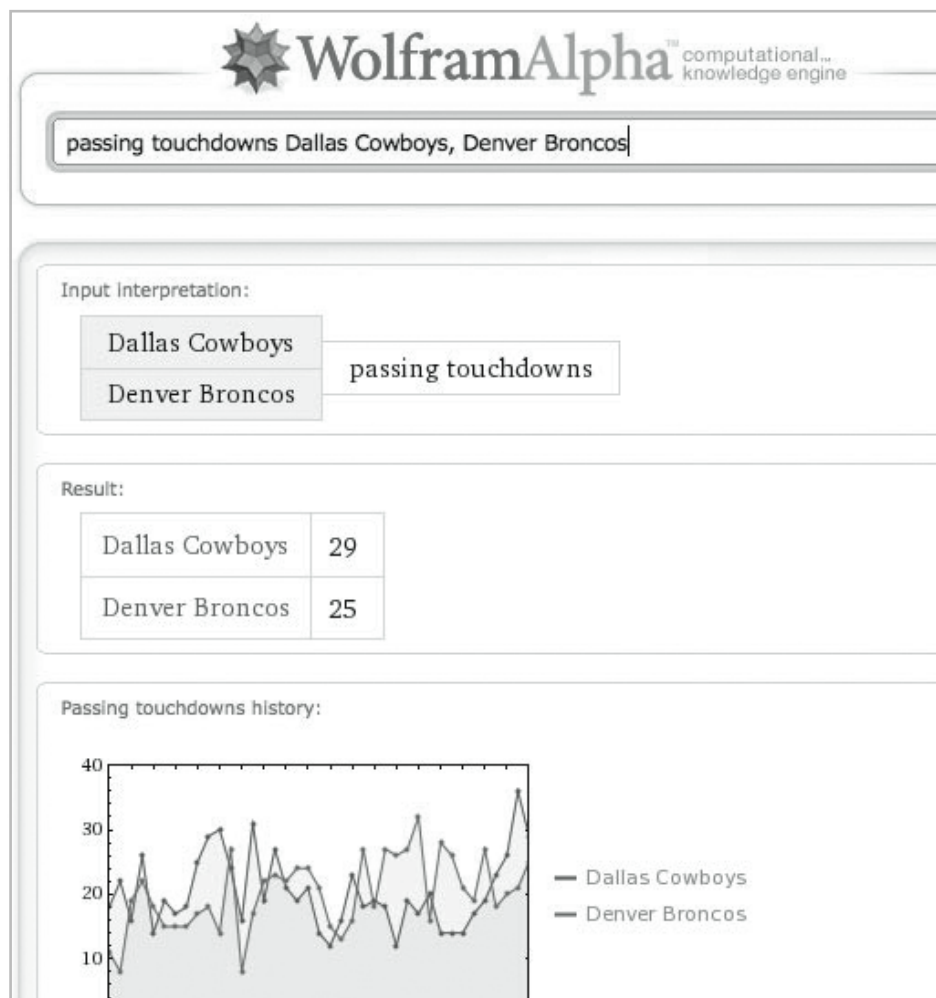
So the web of linked data and the web of documents actually connect in both directions, with links.

User Interfaces for Semantic Content

RWW: At the recent SemTech conference, Tom Tague of Thomson Reuters' Calais project suggested that user interfaces for semantic content are key in getting more take-up. With that in mind, I wonder if you've seen some great

experiment which happened over a particular period, I need to be able to look at that in the calendar - so I can connect the genetics to the calendar.

So one of the things I hope to see is domain-specific things for various different domains, and the generic user interfaces. And hopefully the generic interfaces will be able to tie together all of the domains.



Wolfram Alpha and Natural Language Interfaces

RWW: An interesting new product was launched this year called Wolfram|Alpha, described as a 'computational knowledge engine.' It's kind of a mix between Google (search) and Wikipedia (knowledge), and its key attribute is that enables you to compute something. The founders think that 'computing' things on the fly is something we're going to see a lot of in future. What's your take on Wolfram|Alpha?

TBL: There are two parts to that sort of technology. One of them is a sort of stilted natural language interface. We've seen those sort of natural language queries for years. Boris Katz [from W3C] created a system called START [a software system designed to answer questions that are posed to it in natural language]. I think with the Semantic Web out there, those sorts of interfaces are going to become important, very valuable, because people will be able to ask more complicated things. The search engine has traditionally been limited to just a phrase, but some of the search engines are now starting to realize that if they put data behind them and have computation engines, then you can ask things like 'what's this many pounds in dollars?' and so on. So

yes, those interfaces will become important.

Conversational interfaces have always been a really interesting avenue. We've had voice browser work in W3C, that has been an interesting alternative avenue. It's possible that as compute power goes up, we'll see a proliferation of machines capable of doing voice. It'll move from the mainframe to being able to run on a laptop or your phone. As that happens, we'll get actual voice recognition and pattern natural language at the front end. That will perhaps be an important part of the Semantic Web.

We talked before about what a great challenge the Semantic Web is going to be from a user interface point of view. Conversational interfaces are going to be part of [solving] that. Of course it's also going to be really valuable to have compositional interfaces - for the visually impaired and so on.

Wolfram|Alpha is also a large curated database of data sets. Obviously I'm interested in the big data set which is out there, which is Linked Data. This everybody can connect to. I don't really know a lot about the internals of Wolfram|Alpha's data set. I don't know whether they're likely to put any of it out on the web as Linked Data - that might be an interesting

addition. I imagine that quite a lot of it may have come from the web of Linked Data.

e-Commerce and Linked Data

RWW: There have been reports recently that both Google and Yahoo will be supporting the Good Relations ontology and linked data for e-commerce. Companies such as Best Buy are already putting out product information in RDFa. What would be your advice to e-commerce vendors right now, to help them transition to this world of structured data on the Web. The same question could be asked across many verticals, but e-commerce seems like one area which has some momentum right now. Would you advise them just to put out their data as Linked Data?

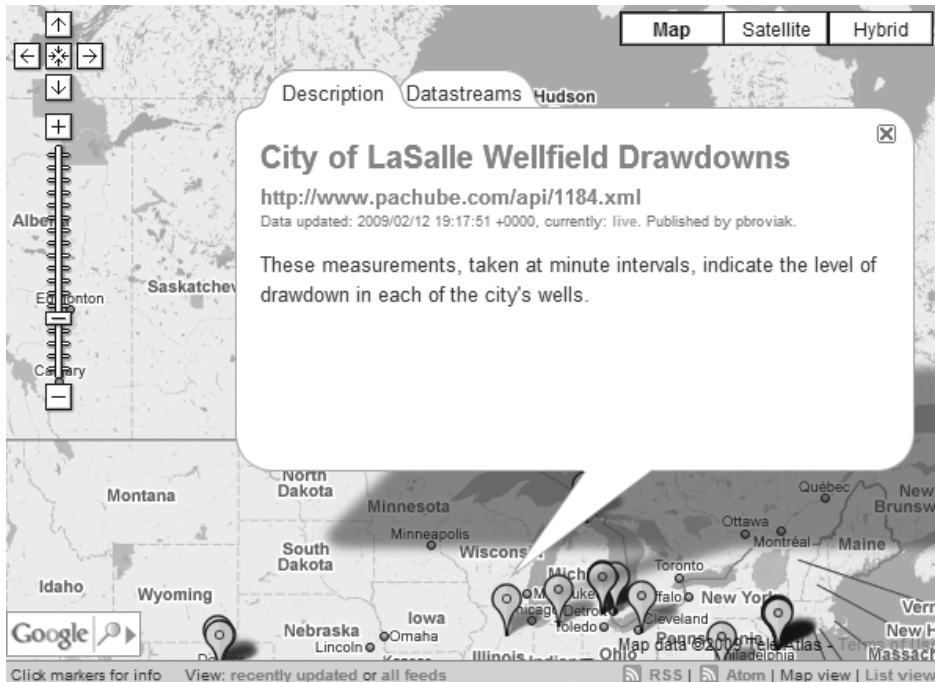
TBL: Yup! Certainly this year is the year to do it. I've been advising governments to do it and when you look at an enterprise, you find that a lot of the issues are the same. But when you put your data from government or enterprise out there, make sure you don't disturb existing ecosystems. Don't threaten those systems, because you've spent years building them up.

Maybe there's an analogy with when the Web first started and the first bookshops went online. They were more or less a flyer, saying 'hey we have a great bookshop at 23 Main St, come on down!'. Let's say that a person named Joe owned one of these early online bookshops. If somebody had suggested to Joe that he should put his catalog online, Joe would've felt that that was very proprietary data. And he'd be worried that other bookshops would see where he was weak, so they'd be able to advertise themselves as filling that niche he's weak in.

But when his competitors Fred and Albert put their catalogs online, then Joe can check which books people are browsing at Fred and Albert's websites. So Joe would [finally] be persuaded to put his book catalog up online. But he doesn't put up the prices... until Albert and/or Fred does. And even if catalog and pricing is up there, nobody puts their stock levels online. And there was a period of time when nobody [i.e. online booksellers] had their stock levels up. But people got fed up with ordering stuff that wasn't in stock. So the first book shop to actually tell you about stock levels suddenly was then unbelievably attractive to its customers.

So there's this syndrome of progressive competitive disclosure. This happens when people realize that if you're going to do business with somebody, if you're going to have your partners up and down the supply chain, really it's useful to check the data web - and life goes much more quickly and open.

Best Buy may be what starts the ball rolling [among e-commerce vendors]. Now if I want to look out for what [products are] available,



I can write a program to see what there is. If somebody wants to compete with Best Buy, to my program they'll be invisible unless they can get their data up in RDF. Doesn't matter whether they use RDFa or RDF XML, as long as it maps in a standard fashion to the RDF model, then they will be visible.

The Internet of Things

RWW: I'm fascinated by how the Internet is becoming more and more integrated into the real world. For example the Internet of Things, where everyday objects become Internet connected via sensors. Have you been following this trend closely too, and if so what impact do you think this will have on the Web in say 5 years time?

TBL: It connects very much with Semantic Web [and] with linked data. With Linked Data you've got the ability to give a thing a URI. So I can give a URI to my phone, and I can say that's my phone in Linked Data. And also the company that made it can give a URI to the model of the phone. They can also put online all the specs of the phone, and then I can make a link to say that my phone is an example of that product. So now any system which is dealing with me and has access to that data will be able to figure out the sorts of things I can do with my phone, which actually is really valuable. Especially if the phone breaks.

The Semantic Web has already given URIs to things, and to types of things. When the things themselves have an RFID chip in them, then I think it's a very exciting world. One can take that RFID chip, go to the Internet and find out the data about the thing. Whether we'll be able to do that, whether the manufacturers will be open enough to allow me to turn data about the

identifier of the thing into data about the thing, is yet to be seen. But it's a very exciting idea.

Similarly, I'd like to be able to scan a barcode and get back nutritional information about what's in - for example - a can of food. But we don't have that yet. To get that sort of thing, which is very powerful, we need to build look-up systems, which allow you to translate an RFID code or a barcode into an HTTP address.

The Semantic Web is a web of things, conceptually. Tying an actual thing down to a part of the web is the last link - the last mile. Give the thing a notion of its own identity in the web.

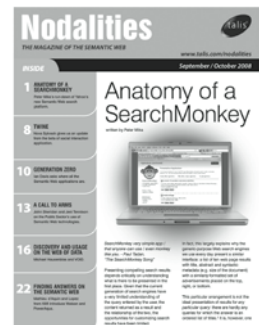
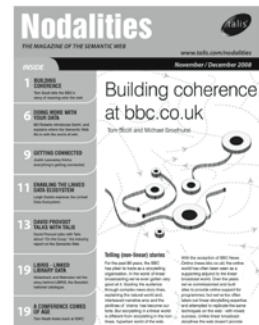
Conclusion

RWW: The over-riding message in both Part 1 and 2 of our interview with Tim Berners-Lee, is for companies and organizations to make their data available online. Preferably as Linked Data, which uses a subset of Semantic Web technologies. But Berners-Lee noted, in Part 1 of our interview, that he'd even be happy with the data in CSV (comma separated values) format.

It's clear that we've seen a lot of progress in linked data already in 2009. In upcoming posts on ReadWriteWeb, we'll continue to track this trend and explain how organizations can contribute their data.

Original article is available at tinyurl.com/tblrww

Nodalities Magazine 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 nodalitiesmagazine@talis.com.

Nodalities Blog

From Semantic Web to
Web of Data

blogs.talis.com/nodalities/

Garlik Announce Open-Source of 4Store

Podcast with Paul Miller



Paul Miller: Hello and welcome to this podcast, with your host Paul Miller. Today I talk with Tom Ilube and Steve Harris from Garlik. We discussed their decision to open source their RDF triple store, 4store, and make it freely available for download.

For more information on this show, please see the notes on the website. Thank you.

Paul: Tom, Steve, thank you very much for joining me for this podcast today. Before we talk about Garlik and open source triple stores and all the other things we're going to get to, I think it would be worthwhile to have you both briefly introduce yourselves. So first off, Tom. Who are you? And what do you do?

Tom Ilube: Sure. I'm Tom Ilube. I'm chief executive of Garlik and founder of the company. I've been in technology for 25 years and was previously chief information officer of one of the world's largest Internet banks.

Paul: OK, good, and that's Egg for those who are listening along and anyone who wants to get their Egg card out and say, "I've got one of them." [laughs]

OK, good. Thank you. And Steve, Steve Harris. Hello. Who are you, and what do you do?

Steve Harris: Hi. I'm Garlik's head of architecture, and I've designed RDF systems and some of our systems as well. And I'm proud about—I was a researcher at the University of Southampton working in the IAM Group.

Paul: OK, good. Thank you, Steve. And we'll dig into some of the detail of that RDF stuff in a moment. First through, you both work at Garlik. Tom, what does Garlik do?

Garlik is an identity protection and fraud prevention company aimed at consumers. We help consumers to look after their personal information in the online world.

Tom: Garlik is an identity protection and fraud prevention company aimed at consumers. We help consumers to look after their personal information in the online world. And we set the company up about three years ago, working with and for consumers as we saw huge amounts of personal information appearing in the online world, and consumers being taken advantage of, particularly by identity theft and so forth.

So Garlik helps consumers look after their personal information.

Paul: And how does it do that?

We created a service called Data Patrol which essentially patrols or harvests and looks for information on your behalf across millions of web pages of dark markets where information is traded.

Tom: We created a service called Data Patrol which essentially patrols or harvests and looks for information on your behalf across millions of web pages of dark markets where information is traded, on structured databases that have been made publicly available, and analyses that information to see whether there's anything that puts you particularly at risk of identity theft or exposing your privacy in a way you didn't intend.

And then it alerts you and advises you on what you should do about it. It's a continuous scanning, monitoring process. We make it available to people both direct and also through business partners such as banks, credit card companies, Internet service providers, and so forth.

Paul: OK, and when we last I spoke, I think because of the nature of the data, this service was just available in the UK. Is that still the case?

Tom: It's now available in the UK and in the US. We launched it in the US towards the back end of last year—set up an office in New York. We've both launched it there, and we're talking to a number of partners. And we hope to be launching in a couple of other European countries—certainly one, possibly two—later on this year.

Paul: OK, good. Despite the fact that this data comes from a lot of different databases, a lot of different websites, and comes out in a lot of different forms, you're actually storing it all in RDF, aren't you?

Tom: Part of the reason we went down the path of semantic technologies was because we knew that in order to deliver the type of service we wanted to, we were going to be wrestling with information from a huge range of sources and coming at us in different formats. Some of these sources of data are quite transient as well. So they come up very quickly. We have to analyse them, understand how to interact with them, and then six months later they've disappeared again.

And one of the challenges, in fact probably the biggest challenge that I've found as chief information officer of Egg for a number of years—and I note that other CIOs find—is being able to change your databases, being able to wrestle with heterogeneous data sources that change constantly it seems. It's a huge problem in large organisations.

What I've found is the types of technologies we're working with now gives us a real advantage in that area. It gives us a degree of flexibility and adaptability that I just didn't have working with the previous set of technologies.

What I've found is the types of technologies we're working with now gives us a real advantage in that area. It gives us a degree of flexibility and adaptability that I just didn't have working with the previous set of technologies.

Paul: OK, and this underlying data storage is entirely homegrown, or have you brought things in from outside?

Tom: The RDF storage is entirely homegrown. Do you want to comment on that, Steve?

Steve: Yes. We use Dave Beckett, who works for Yahoo!, we use his SPARQL and RDF parser, but the actual storage and the execution engine are all completely homegrown.

Paul: OK. I guess that is sort of the meat of the story today, is that you're taking the step of taking this system that you've developed over a number of years, presumably at a fair degree of expense, and you're making it freely available for everyone to have. Are you mad?

Tom: [laughs] I was almost going to say absolutely, then I heard the question at the end. "Are you mad?"

Actually, we've been thinking about this for quite a while and talking to our investors and so forth. And we decided take our RDF store, which we call 4store, and make it freely available to anyone anywhere—businesses, governments,

developed at Egg. Maybe some of those could have been released as well.

Paul: One of the things in the press release, you mentioned that Garlik may be willing to provide sort of support to help other companies get going with this. Is that a new business direction for Garlik? Are you going to do that for free as well?

Tom: We won't be able to do that for free, unfortunately, but what we felt is it's useful to some organisations to know there is a company that's willing to stand behind the software and provide some support and guidance as

want to do this. Here are the reasons why. Is this something you'd be willing to support us in doing?"

And it was really only about a week or so ago that we had the final board meeting to discuss it and decide that we were happy to do it, and we had their blessing, and let's get on with it. Actually, they're quite excited about it.

Paul: OK, good. So the desire to do this would have been very clearly on the table in your last funding round, for example?

Tom: Absolutely. We've talked to our VCs over the last year, and we did our last funding round about six months ago—maybe less than that now. So this has been in the conversation all along.

Paul: OK, good. Now. Steve, one of the things that always comes up as soon as we start talking about triple stores is how big will it go. So for 4store, how big will it go?

Steve: That question is always a bit hard to answer because it's always more a case of, how fast do you want it to be than hard limits. But the production system that backs Data Patrol has currently about 15 billion triples, long-five. And that runs fast enough to supply that service to not many users.

It's a little difficult to predict, because it depends on the shape of the data and how complex your data is and the kind of queries you want to run.

So it really depends how much hardware you throw at it, and how fast you want it to be. It's a little difficult to predict, because it depends on the shape of the data and how complex your data is and the kind of queries you want to run.

Paul: OK, but in the press release there's the figure of 60. 60 billion is mentioned. Is that a projection, presumably, of how far you think it could go?

Steve: That's based on partly projection, and partly based on non-production test unit we've done.

Paul: OK. And that makes it pretty big actually. I mean, if you look there's a page on the World Wide Web Consortium site that I'll link to in the show notes that talks about large triple stores. And most of them are sort of 10, 12, 15 billion. You're bigger than that already, and project you could go three or four times that size.

The reason why we wanted to do this was because as a company we placed quite a big bet. We are betting that this set of technology—RDF, the semantic web, the web of linked data—are going to become hugely important.

universities, academics, developers—as open source software. And we're making it available today for people to download and really do whatever they want with it free of charge.

The reason why we wanted to do this was because as a company we placed quite a big bet. We are betting that this set of technology—RDF, the semantic web, the web of linked data—are going to become hugely important. And really, as we look forward, are going to define what the web looks like in many years to come.

And therefore as a company it's actually in our interest to help make that happen. So if we have technology that can really help other people to store and publish data in RDF form, then that's what we want to be able to do. And so once there was some computations to be had along the way to get everyone aligned, the idea of doing, from our company's point of view makes a lot of sense.

Paul: But you're not a software company. You're a company solving identity management problems that happens to need to use software to get there. Is this not a dilution of your core mission?

Tom: No, it isn't really. One of the advantages, in a way, is we've built this software to be used for our own purposes and it's been in production for the last three years. So it's a good, hardened bit of software and I think people will see that when they try and download it, install it, and make use of it.

It's not a distraction for us at all to release it like this. In fact, I think more companies ought to be looking at their software portfolio and thinking about what they can release. And perhaps with hindsight, I could have looked at some of the great bits of software that we

required. Therefore we decided that we would do that.

We're not trying to build up a whole major business line in that area. We're not transforming into a Red Hat or one of the other organisations that supports open source software. But we have developed the software. We understand it inside out.

I'm sure there'll be a lot of people quite happy downloading it, installing it, and getting on with it without any support. And there'll be some organisations that'll want to be able to turn to us and ask for support. And we'll be there to provide that.

Paul: OK. And you're venture capital backed—Garlik.

Tom: That's right.

Paul: How did you get his past VCs?

Tom: [laughs] The one thing I would say is if you're going to do something like this, you don't spring it on your VCs. It's sort of a long process. So we probably started talking about potentially doing something like this maybe a year or so ago with our VCs. Amongst ourselves, we've probably been talking about it earlier than that, but with our VCs perhaps a year ago I would of started to raise it at board meetings.

To say, "Look I think we might want to do something like this." But I wouldn't push it too far straightaway. What I was essentially doing was seeing whether my VCs were aware of these sorts of approaches, what their gut reaction was, and so forth.

Essentially, we've taken them on a journey. Particularly over the last six months it's sort of accelerated, where we've said, "Look, we do

Steve: Yes. Scalability was one of the major design features. We knew we were going to have a lot of data to support and service. So from the outset we accepted some of those scale problems. 4store not only works at that size, it's also still quite fast. If you have a sufficiently powerful machine, you could import data at 120 or 150 thousand triples a second all the way up to tens of billions size.

Scalability was one of the major design features.

Paul: So to achieve these scales, have you done something fundamentally different to the others? Or have you simply optimized existing practices better than they have?

Steve: Yes. It's built from the ground up to be a clustered system. Although it does run on a single machine, a cluster of one. When you take that approach, it leads you down a different design route.

Another thing we've done is, rather than building it out of a modern more trendy language, we built it out of C just to get that raw performance, especially in the IL area. We found it very hard to get the same amount of performance out of higher level languages.

Paul: OK, and you talked about designing it for clustering. Is 4store designed for high performance machines in a cluster, or is it designed to run on commodity hardware?

Steve: We run it on commodity one looped Linux service. It does like a lot of RAM, because it tries to put as much of the index in RAM as possible. But really it's not all the fussy. Other people that are running it are running it on smaller clusters of much more powerful machines. It self-adapts to whatever hardware environment it's installed on.

Paul: OK. And we've been talking quite a lot about downloading it, and clearly it's available on the website now to download. Do you have any plans to offer a hosted arrangement for people who don't feel capable or who don't want to download and run software of their own?

Tom: We're not really planning to go in that direction. I think that there are a number of companies that do that, and I suspect they might be interested in taking our software and providing that sort of hosted arrangement. But it's not sort of a service line that Garlik itself will get into.

Paul: OK. And what sort of machines will it run on?

Steve: Internally, we've only ever run it on Red Hat derived Linux machines and on Mac desktop machines for our developers, just to use single machines. It's a fairly portable piece of software. The clustering support depends on a multicast DNS library which is not all that portable, but other than that it should run on pretty much any Unix system.

The only thing we've really got any experience running it on is Red Hat derived systems. We tend to use Centos.

Paul: OK. And presumably people can download it for Ubuntu or whatever it may be and test it. And if it doesn't work, find out why it doesn't work and contribute those fixes back to the code line.

Steve: Yeah, I think because it runs on OS10 and on Red Hat, it wouldn't have any problems on something like Ubuntu. I think if you wanted to try and run it on, say, Open Solaris you might find a couple of challenges, but I think any Linux system would be very straightforward.

Paul: OK, good. So you've intrigued me. You've got me interested in this massive triple store. I've got a big data problem in my office. What do I have to do to get going once I've downloaded it? I'm probably use to Oracle or MySQL or something. Presumably it's quite a shift to move from thinking about a database to thinking about an RDF triple store.

Rather than building it out of a modern more trendy language, we built it out of C just to get that raw performance, especially in the IL area.

Steve: Well, not really. We designed it so once it's installed it's actually very easy to use. It doesn't have exactly the same model as an Oracle or a MySQL. If you've got a cluster set up or a single machine, you start an individual backend for every database that you want to query.

And then you just run an HTTP server, which attaches to the database itself to provide the SPARQL endpoint. And once you've done that, you can use RESTful HTTP operations to add and remove data. Or there's command line tools if you prefer that kind of thing.

Once it's up and running, it's very simple to use.

Paul: OK, and you said RESTful APIs for pushing the data in and out? So that's how I would attach it to my existing applications, is it?

Steve: Yes, that's how we use it in our... So the QDOS QDOS.com services and also Data Patrol use the RESTful API on top of their SPARQL endpoints to add and remove data.

Paul: OK. And how fast will that work?

Steve: So for bulk imports, it's slower to put individual HTTP RDF files over HTTP than it is to use the command line tools. Again, it's really difficult to tell. It depends on the exact structure of the file. But for sort of normal sized RDF files, it's not really a problem—40 to 50 thousand triples a second should be easily attainable.

Once it's up and running, it's very simple to use.

Paul: OK. Good. Now one of the domains obviously where there's quite a lot of interest in semantic web stuff at the moment is the whole area of government. We were talking before the recorder went on about some of the work that Tim Berners-Lee and others are doing in that space at the moment. It seems to be attracting quite a lot of attention from outside the semantic web community.

Is something like 4store—are we going to see it running in the home office any time soon?

Tom: I think it will be up to the various government departments. If they find it interesting or not, it's available to anyone who wants to use it.

I think there are a few different stores and open source RDF stores out there. Yeah, some of them are sort of more feature-rich and perhaps slower, and ours is fast and large and basic. And I think different organisations will decide what they're requirements are, what they're trying to do, and decide whether ours lends itself to what they want to do.

But I think overall we would be happy. The more organisations that are making raw data available in RDF form, the better. So if our store contributes to some organisations doing that, then that's a jolly good thing.

Paul: Absolutely. "Raw data now," as Tim Berners-Lee has been saying recently.

So the store itself will allow an organisation, whatever they may be, to clearly use that as the backend to some application of their own. Is it also capable of contributing the data that's in the store back out to this wider linked data cloud? Or does it simply sit behind applications?

Steve: So the QDOS.com service is also a linked data application. And the linked data interfaced to that is provided by 4store. When

a linked data press comes in, it's just rewritten into a SPARQL construct statement. And that's very fast. It's quite highly optimised.

The SPARQL interface was designed so that it could be exposed to the outside world. It's got a lot of features in there to—some of the SPARQL features disabled by default to make it safe to deploy inside enterprise environments. A typical SPARQL system is not safe to be deployed inside a secure enterprise system because of some of the features of SPARQL. So that stuff's disabled by default.

And also it has some features in there that enable you to limit the amount of effort you're willing to put into arbitrating any particular query, dependent on where it's coming from. But we've never done that on Angus so we're not prepared to say it's safe to expose the SPARQL endpoint and let random people on the street come up and try a SPARQL query at the gates. For that, for certain, you want to design criteria.

Paul: OK, so the intention there would be that if I or someone else either deliberately or accidentally sent a very expensive query to your

Paul: OK. Now you're making this available under an open source license. So presumably people are going to take it and they're going to start tinkering with it, and they're going to start adding things in that they feel they need and want. Are you going to contribute those additions back into the code line?

Tom: Yes, absolutely. Yes. So our needs inside Garlik in some ways are quite better. Although we have a lot of data, we don't do things like natural language search and so on, although it wouldn't be too hard to add that kind of thing to the engine.

So if people have those needs and they're willing to make those patches available, we'll definitely include them in the main line.

Paul: OK. And do you think continued development will make it get bigger as well, or would you almost have to start again to build something that was capable of scaling further?

Steve: Well there are two things on that really. It could probably improve and get bigger in certain respects, as it is. But what we've done internally—we've been working on building

But I think people will find they can go a long way in the direction they want to with 4store and other stores that are out there in similar sort of scale.

Paul: Yeah, good. So are you going to be actively and aggressively pushing 4store, or is it simply going to be put out there? Built it and they will come?

Tom: We'll certainly want people to know about it, so I think we would really like it if people were thinking about using an RDF store and having a look at what was out there and what was out there in the open source world, then we'd hope that 4store would be one of the things that they look at.

But I don't think we're going to be sort of trying to compete with others, in a sense. I think that the more options there are for new organisations and new people looking at the whole linked data world, the better.

I know when I was a CIO, if I was trying to make a technology decision, I always liked it to be three or four or five options with different characteristics so that I could decide what was right for my organisation. So we're keen that as people are looking at making this sort of decision of moving in this direction, this is one of the options that they look at.

And we would really like to see the more organisations that are moving in this direction that are using tools like ours and others to publish their data, the better. Because as a company what we're excited about is what the world looks like in three or four years time when thousands and thousands of mainstream organisations are publishing their data as RDF, part of the Linked data cloud.

That will be a fascinating world to play in. And we're hoping that what we have done is to help to contribute towards that.

The store itself will allow an organisation, whatever they may be, to clearly use that as the backend to some application of their own.

system, it should be able to say, after a certain period of time, that it's going to stop and fill the query.

Steve: Yeah, it does IL accounting, so it'll allocate out a certain amount of effort. And you can say how much effort you're willing to expend on a query. And then you get only a partial result and a warning saying that it's partial, and that it was limited by the complexity.

Paul: Right. And presumably then if you're a valid query, you can contact the system provider and ask them to give you more time to run your query.

Steve: Yeah, you have to provide a separate endpoint that has slightly more resources allocated to it.

Paul: Right, OK. I guess one of the more real-world applications of that would be perhaps that internal systems get more time than external systems do.

Steve: Yes. That's what we've done. We have provided SPARQL endpoints to a few people that want to run applications on top of QDOS, but our internal ones go through a different endpoint with a much higher limit on it.

another generation of RDF store. The scale is probably on order of magnitude bigger than 4store does.

And as their systems and our internal needs grow, then we'll start to use that store to underpin them. And that one will scale much, much, much larger than 4store will.

Paul: And is this called 5store?

[laughter]

Steve: Internally we call it Keep, but we haven't got an external name for it yet.

Paul: OK. And presumably there's also this sort of related school of thought that suggests we don't necessarily need to lump everything into one store. So 60 billion triples or there about may actually be big enough for any one store, and we need to be getting better at creating the connections between stores working together.

Steve: Absolutely. Yeah, yeah, and we're thinking about...

It think it'll vary between—there'll be some huge organisations that decide that everything they want to do internally and actually it turns about that they do need big centralised stores.

We would really like it if people were thinking about using an RDF store and having a look at what was out there and what was out there in the open source world.

Paul: What would that world look like? If all these claims certainly for government data become real and if organisations of various other kinds follow along and start exposing data. Wearing your identity hat for a second, I guess, as a corollary perhaps to the brave vision, what would that world look like?

Tom: Gosh, I think the organisation would be a fascinating world. I try not to predict too much what it would look like and just try and help to make it happen and then get stuck into it. But I think the issues of identity and privacy and control and power over your personal information shoot up the agenda hugely in that world.

In a world of masses interlinked data wherein you are and who said what and what personal data there it is very visible, then it becomes a real challenge, but also a real opportunity for individuals to understand what data is out there about them and what they can and can't do with it.

I think my view is that the world of personal information is changing in a really sort of fundamental way. And we're going to have to stop talking about wouldn't it be great if things were how they used to be? And how do we hold back the tide? And how do we get our privacy back in the box, as it were?

And recognise that the world is changing quite fundamentally and the challenge is how do you as an individual survive and thrive and take advantage of opportunities in that world. I think it is sort of two sides of the coin, both protecting yourself and taking advantage of the information that is out there about you and others.

Sort of two sides of the same coin. We're going to have some really interesting times over the next few years, thinking about how you live in that world.

Paul: Indeed we are. And it's going to be an interesting journey and interesting finding out how we cope and how we make the most of it because, as you say, it is a big opportunity for all of us and for organisations building systems and technologies to take advantage of it.

So it's going to be an interesting time. And as you say, 4Store is there on the table for people who want to start looking at how they can make use of all this data. Now, within the world of triple stores, there are an awful lot of them.

Most tend to be an awful lot smaller. Do you think we are going to see some harmonisation in that market? I'm not suggesting that it will be just one. But do you think we are going to see some of the smaller, older ones falling by the wayside?

Steve: I guess there will be some that fall out of use, but if you look at the relational database market, which is the nearest equivalent. There's really an awful lot of software in that domain, just spanning different needs and different emphasis and catering different languages and so on.

So I think the number of RDF stores we've got at the minute certainly isn't unsustainable.

Paul: OK. And presumably if people keep using them then there is still a need for them.

Steve: Yeah, exactly. Yeah.

So it's going to be an interesting time. And as you say, 4Store is there on the table for people who want to start looking at how they can make use of all this data. Now, within the world of triple stores, there are an awful lot of them.

Paul: OK. Good. So just before we finish off then, do either of you have any closing thoughts on where we go from there?

Tom: I think that what we need to do, those of us that are close to this area, is to try and bring the wider technology world into it and to make them aware of the advantages of adopting these technology sets and at least exploring them and trying to get their minds around them.

There are hundreds of thousands of organisations out there that will need to be looking at these sorts of technology sets, from us, from other people, thinking about what it means to their businesses. And I think one of our challenges, if you like, is both to make our toolsets available for those sorts of organisations to engage with without making too much fuss, but to give them a way in which they can start to explore without making a massive commitment.

And also figure out how to speak to the wider world in a language that they understand. I think that sometimes you can get so wrapped up in the world that you live in that you can forget the 90 odd percent of the rest of the world has no idea what you are talking about.

I think sometimes we run into that problem in semantic technology world as well. But I think over the next sort of couple of years, getting more and more organisations, more and more smart technology people who just haven't happened to encounter this world particularly to come and play and get involved and experiment and try things. It's what we need to do.

Paul: Yeah. Steve?

Steve: No, what Tom said, really. I think getting the message out there that this real world practical problems that can be solved quickly and efficiently and sustainably. Technology is

the key thing. That is the main benefit we see is the flexibility and the ease of use of relational databases.

Paul: And you actually have real world examples to point to as well, because you have been using it in Angus yourselves for a number of years?

Steve: Absolutely, it makes a real difference being able to say to organisations—I've been a CIO myself. You can be quite nervous. You are thinking about introducing a new set of technology and it looks like there's only downsides. [laughs]

It looks like if you introduce this new technology your chief executive or your marketing director is going to come and slap you around the head and say, what on earth did you do that for? And the old stuff was OK, wasn't it?

There are hundreds of thousands of organisations out there that will need to be looking at these sorts of technology sets, from us, from other people, thinking about what it means to their businesses.

So before you introduce something new you need to be pretty convinced that it is going to give you some real benefits. And if you can talk to someone who has been using it to support tens of thousands of users in a production system day to day, even if you don't use that same technology set that they are using, the fact that there are people out there like us and others using this technology in production day in and day out, to support large scale and growing businesses. It means a lot to the main stream technology guys.

Paul: Yeah, definitely. Good, well thank you both very much for that. and I look forward to seeing what people make of 4store. They can go to your website and download it and try it for themselves.

So thank you and good luck.

Steve: Thank you very much.

Tom: Thanks.

Nodalities Magazine 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.*



Read Previous Issues of Nodalities Magazine Online

Want to contribute? Please visit www.talis.com/nodalities or email nodalities-magazine@talis.com.

[About](#) | [Contact](#) |

[About Nodalities](#) | [Previous issues](#) | [Nodalities Blog](#)

Nodalities


THE MAGAZINE OF THE SEMANTIC WEB



Contact us

+44 (0) 870 400 5000
Email editorial

Previous issues



Nodalities Magazine

Semantic Web technologies are moving beyond the laboratory and finding a role in the wider business world. Traditional companies such as Reuters and Oracle are putting semantic technologies to work, and startups are attracting increasing venture capital investment.

Talis has launched a magazine called Nodalities that bridges the divide between those building the Semantic Web and those interested in applying it to their business requirements.

Want to subscribe?

[Subscribe for free](#)

Want to contribute?

[Share your ideas for future articles](#)

Updates

[Follow us on](#)



ISWC **2009**



**WESTFIELDS
CONFERENCE CENTER
FAIRFAX, VIRGINIA USA**

8TH INTERNATIONAL SEMANTIC WEB CONFERENCE

25-29 OCTOBER 2009

Call for Papers and Hotel Reservations Now Open, Visit:

ISWC2009.semanticweb.org