Nodalities

THE MAGAZINE OF THE SEMANTIC WEB

talis®

www.talis.com/nodalities

Issue 8

INSIDE

Read Previous Issues of Nodalities Magazine Online

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

twitter

Follow Us On Twitter
www.twitter.com/nodalities

GraphOS: the data web as an operating system

Alan Morrison, Editor of PWC Technology Forecast

Since its inception, the Web has been consistently teaching us how to do business better. It taught us more efficient ways of document sharing and management. It taught us to create and serve up audio and video in surprisingly handy and egalitarian ways. It taught us how to buy and sell quickly online and in the process take advantage of what we now call the prosumer. And it taught us how to connect with customers, crowds of developers and each other in a many-to-many fashion that outpaces traditional broadcast and one-to-one methods.

We haven't fully experienced the ramifications of these disruptions, but the fact is that the Web's not done teaching us new things. The most powerful of these might be data that only need to be created once, and then used anywhere, anytime, by anyone. A data architecture that supported this capability would be the foundation of a true Web operating system, one that would favor global applications that serve the data, rather than the other way around. With a Web OS like this, we'd have a system designed from

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

Editor's Notes

Welcome to the eighth edition of Nodalities Magazine. Throughout the past couple years, we've seen the Semantic Web move from a great idea through being an innovators' dream into proof-of-concept phase. We've explored many of the changes and movements here in this very magazine, and I'm very grateful to have had such positive feedback on Nodalities and our contributors.

As always, we make this publication available free of charge in either print or online form, and we'll ship world-wide. If you'd like to receive a print or online subscription, just sign up on talis.com/nodalities.

As the world learns to cope with its increasing creation—and reliance upon—information, we've seen many organisations and commentators working on predicting the next advances and strategies. To us, naturally, it seems that the infrastructure of the Web and the innovations of linking data will help ease this burden and lead to efficiencies and further innovations.

In this issue, Alan Morrison from PricewaterhouseCoopers' Tech Forecast gives us an overview of the world of interconnected data from his business perspective. He presents the idea that the world's data could be put to good use if only they were not inaccessible to the growing compute power being generated. Paul Miller also talks about the related area of cloud computing and explores the potential environmental impact of all these farmed computing facilities running all the time.

What is DataIncubator? What data are being incubated? Leigh Dodds introduces the project and discusses some of the data growing in its care. Michael Hausenblas and Richard Cyganiak from DERI in Ireland give us a look at their remixing of the Web of Data through Sig.ma: their Linked Data browser. Talis tech evangelist Richard Wallis takes us through the jungle of pre-existing applications and how they can make use of Juice: a JQuery partner to Linked Data.

Finally, MySociety's Tom Steinberg talks with Talis about his work with public information in the UK.

If you have any feedback, suggestions or you'd like to contribute to a future edition of Nodalities Magazine, please feel free to get in touch at zach.beauvais@talis.com.

Continued from front page.

the start not to orphan data. That system could be quite powerful inside governments, where the data that get created in one set of applications owned by one agency sit in silos and compartments inaccessible to another agency. This is not to mention application data and documents that the public at large finds inaccessible or hard to take advantage of. Today, we're trying to build bridges to alleviate this problem, but that can't be the long-term answer.

Today, applications have to be linked more or less on a one-to-one basis to talk to one another. Repositories are separate. To maintain separate repositories implies inefficiency and a lack of scalability.

Today, applications have to be linked more or less on a one-to-one basis to talk to one another. Repositories are separate. To maintain separate repositories implies inefficiency and a lack of scalability.

Instead of designing in unnecessary redundancy, we could have application data designed from the get-go to be a part of a larger, queryable whole. This way, not only would we reduce redundancy and create a data scalability, we'd amplify the power of each application. The self-describing data providing the foundations for these applications would be truly virtual as far as the applications generating them are concerned. Applications would have a data foundation that's inclusive and global. In the process, the applications themselves would become less redundant. Enterprises who now face the daunting prospect of how to rationalize thousands of applications couldn't help but find this kind of architecture appealing.

The Need for Many-to-Many Data

For this objective to be possible, machines on their own need to be able to find their data and connect them to other data, so that applications can access them as a virtual whole. A data model that allows global identifiers compatible with those already in broad use for pages on the Web—but more specific—would be essential, as well as a way to enable standard n-dimensional relationships between individual data elements. The problem of orphaned data—data that are created only to sit in silos, where they often end up forgotten and underused—is among the most nagging IT problems we have, and yet it's among the most overlooked because it's a systemic problem that's difficult to fix.

A logical approach to this problem would be to fully expose each data element in a standardised form. But this form would need to maximize the simplicity and flexibility of interconnection between elements—something traditional relational databases don't do very well. What we need is a move from a data-base system to a huge graph. Basically, graphs like these employ a Tinker Toy-style approach where each hub has the potential for numerous different connections. Once you have the graph structure, it then becomes a matter of identifying and describing each data element sufficiently so that any machine can locate it. Each datum or bit of data would begin with one context associated with one usage. Other users could then add additional contexts for other kinds of usage. Once the data are sufficiently describable in this way, they naturally become a part of the data cloud and broadly useful with the benefits associated with the network effect.

If we describe data in a standard way to begin with, we only need to create the elements once. What this really implies, though, is creating a small bit of a data-web ready-made for multiple contexts that can be added over time. Automating the procedure of adding context is certainly possible now and, over time, we'll be able to add context as necessary for different usages in a scalable way.

Posting the context and the data where others can get to them and modify them in wiki style would help tackle the scalability problem. The result would be data at Web scale, an extensible data fabric, rather than a series of silos. There will be reasons you would still want to duplicate data—say, for the purposes of pure transaction speed or access control—but if you don't have these special requirements, you won't need to duplicate nearly as often. Even then, you could still rely on one posting as a primary, ultimate source.

Like other things on the Web, the data graphs that are most useful would be the most linked to, the most reliable, dynamic and frequently updated. We'd be able to construct a data meritocracy not unlike the document meritocracy we have on today's Web.

How the data themselves become sufficiently networkable and extensible from cradle to grave becomes a key challenge, but it's a challenge the Semantic Web community's been working on for years. The currently proposed starting point is the Linked Data initiative. Closely associated is the Resource Description Framework, or RDF, which provides the simplicity, connectedness and flexibility of a graph alongside an initial skeleton of meaning that data descriptions—as simple as folksonomies or as expressive as RDF Schema and Ontology Web Language (OWL)—can latch onto. Then ideally, the identifiers and data descriptions should themselves be networked along with the data.

In this scheme data, identifiers and descriptions are part and parcel of the same atomic structure, and all are out on the open Web. They're addressable by anyone to be repurposed, updated or improved. Like other things on the Web, the data graphs that are most useful would be the most linked to, the most reliable, dynamic and frequently updated. We'd be able to construct a data meritocracy not unlike the document meritocracy we have on today's Web.

A True Web OS

A Web operating system that's worthy of the name, one that really does reach across silos and continents of information, could be built on this kind of data foundation. An operating system like this would make applications aware of and able to borrow from each other. Rather than just a Webtop with a browser-based office suite, a true Web operating system would weave, use and repurpose an ever expanding data fabric. It would be data driven, and it would assume the ability to query across data boundaries.

Operating systems in the traditional sense of the term provide the interface between hardware and applications. Data in this scheme aren't even considered as a cohesive layer of their own—they're depicted as individual cylinders, each of which belongs to an individual application. For the Internet and the Web to form a seamless abstraction layer we can build on most effectively, a seamless data underlay is essential. Data-driven applications would make better use of what's already available in repositories; they'd be designed to take advantage of a continuous data fabric, rather than just individual bolts of data cloth.

A data-driven approach would redirect the information flow and stem the tide of disconnected application proliferation that to this day goes on and on. Instead of just talking, data-driven Web applications would able to listen too, and not just to users in the conventionally interactive sense. They could listen to other applications. One application could create data that facilitate interactions with another application, and make use of functions that already exist, again avoiding redundancy. They could also browse and make use of the relevant data they find that already exist out on the Web.

If there's a sector of the global economy that would benefit most from an elimination of orphaned, redundant data, it's government.

We're just now realizing that for nearly seventy years, since the inception of electronic computers, we've had the cart before the horse. Applications should be designed to serve the data, not vice versa. This is one thing we haven't thought about much yet. A good open Web OS and all its associated applications would have a fully networked data model in mind from the start, with a data identification and description capability that accommodates both the spare and simple and the rich and complex. It would involve everyday users in the data description process because they help provide the context and are the source of the demand.

At the core of the capabilities we need to nurture for this Web OS is the mechanism for describing the data, a Web-centric system for metadata that can accommodate a range of data description types, from simple tags to ontologies. For this system to work, putting the metadata out in the open in a form that everyday folks can help to manage is critical.

A Bigger Vision for Open Data

We already have the basic components of such a metadata system in place with the schema and ontology methods laid out the World Wide Web Consortium's (W3C's) Semantic Web standards. The philosophy I've just described behind these components is nothing new; the ideas have been around since the 1990s and elaborated for over seven years now.

The challenge the Semantic Web community hasn't really fully taken on yet is to make the system more accessible and give people a reason to use it. To get more garden-variety users access, they need an assisted way to contribute to the richness of the metadata. Semantic wikis and autotagging tools are a step in the accessibility direction. In the "reason to use it" direction, there are Semantic Web-based

applications and data stores that are gaining popularity.

But these pieces should be part of a bigger vision. Taking the Linked Data initiative and moving it into an Open Web OS direction might help get more developers on board; many of whom seem to be either oblivious to or skeptical about the need for fully networked data and a data-driven applications architecture. Even more promising are the semantically inclined knowledge workers, as well as the tinkerers behind scripting, mashups, spreadsheet macros and the like. They don't have the biases of the well-trained and indoctrinated who sometimes can't get their thinking beyond relational data.

Purists turn their noses up at those who don't have formal programming backgrounds, but the truth is that programming itself as it progresses is becoming less of a dark art and more of a user-friendly cluster of languages, frameworks facilitated by extensively

configurable user interfaces. The end result is that more ordinary folks are becoming programmers of a sort. We should invite this—it's a way to get business units fully engaged in IT.

What these budding apps enthusiasts need is an awareness of how to do data right. Somehow, a critical mass of those who are knowledgeable about properly networked or "linked" data needs to emerge. The Semantic Web community needs to lead, not follow with its data architecture, precisely because that architecture is so fundamentally disruptive.

To succeed, it needs to articulate what a data-driven application environment consists of and what the other components of an open Web OS would look like. It needs examples of what the apps could do. Once the most curious and experimental get excited about how an open Web OS could turn the apps world upside down or at least sideways, we'd have some momentum.

> *To accelerate the development of an open Web OS, we could emulate the way the Global Positioning System (GPS) was developed.*

To accelerate the development of an open Web OS, we could emulate the way the Global Positioning System (GPS) was developed. The open data initiatives that are springing up in various governments could remove barriers and raise awareness. A full-blown Semantic Web community/government collaboration would magnify that level of awareness and encourage mass involvement. If there's a sector of the global economy that would benefit most from an elimination of orphaned, redundant data, it's government. The public sector could be the one that leads the rest.

With the right vision and resources, government could get out of the habit of orphaning data and build the Web OS equivalent of a GPS system that the private sector could make use of too. Key to success here is not building bridges to connect what already exists, but rather focusing on implementing an architecture that doesn't create orphaned data in the first place. The Semantic Web community would do well to pitch in and help wherever they can with that architecture, because what they've developed is fundamental to it, with no other equivalent on the horizon.

Alan a senior research fellow at PwC's center for technology and innovation and an editor of the firm'sTechnology Forecast (http://www.pwc.com/techforecast).

DataIncubator: What Is It & What's In It?

By Leigh Dodds, Platform Programme Manager at Talis

The Linking Open Data project has had a huge amount of success in bootstrapping the burgeoning Linked Data cloud. There's now a definite sense of momentum behind the project, and a growing number of organisations are now seriously investigating how their data could further enrich the growing Semantic Web, and how the underlying technologies may help them to innovate and explore new opportunities.

The Linked Data community has rightly begun to look at the next round of challenges: What can we do with all this data? How can it be pressed into service to create new applications? What kinds of frameworks do we need to support consumption of Linked Data? But it is important that we shouldn't lose sight of the fact that there's still a huge amount of evangelism to be done and a great deal of data that could and should be part of the web of data. The Linked Data landscape is still not fully mapped out. In short, we need to keep up the process of accumulating, converting, publishing and linking data in as many different subject areas and disciplines as possible.

A full-blown Semantic Web community/government collaboration would magnify awareness and encourage mass involvement.

To date, the bootstrapping process has been supported by a number of community lead projects that convert and re-publish datasets to bring them into the web of data. The recently founded DataIncubator project (http://dataincubator.org) aims to adopt this same "show don't tell" approach, but with the addition of some best practices and with an eye on long term sustainability.

Sustainability, Repeatability, Reusability

A key goal of the project is to lightly formalise the way these dataset conversions are carried out to make sure they are sustainable, repeatable, and reusable. But why are these particular aspects important?

Firstly, lets consider sustainability. As usage of the Linked Data cloud grows, we need to make sure that new data being added isn't going to disappear later—e.g. because a small project website goes offline; or because the original project owner loses interest. It is critical that as serious applications begin to be built against this data consumers can rely on it. One of the primary ways the project is ensuring sustainability is through making use of the Talis Connected Commons scheme (http://www.talis.com/cc). All of the public domain datasets that are converted and published through the DataIncubator project site are being hosted in the Talis Platform. This takes full advantage of the free data hosting offered under the Connected Commons initiative. Talis is therefore contributing to the sustainability of that data.

One of the primary ways the project is ensuring sustainability is through making use of the Talis Connected Commons scheme (http://www.talis.com/cc).

The second aspect to consider is repeatability. The first goal is to make sure that the data conversion process is itself repeatable—that is: we can easily re-generate the data to allow for modelling changes, bug fixes, and the ingesting of new data. And not just now when a project is active, but in three years time when the project may be picked up and extended by a number of other contributors. Ensuring that each of the incubated datasets is supported by open source code makes this more achievable. Ideally, the original dataset owners will be convinced by the benefits long before a project goes stale, but it's important to recognise that evangelism can take time and that different industries move at different speeds. There are already a few Linked Data and RDF projects on the web that model and re-publish the same basic dataset in other ways. By trying to build a community around curating the conversion of a dataset and not just the data itself, DataIncubator hopes to avoid these issues.

The final aspect is one that is often over-looked: how can the original dataset owner build on what the community has created? How can the community's efforts by reused? Reusability is enabled by ensuring that the conversion code is open source and that schemas and modelling design decisions are well documented. This can lower the barrier to entry facing data providers or publishers looking to embrace Semantic Web technology. This is the case particularly where the data conversion is acting on source data(e.g. open, but not linked data). In this case, the data owner may merely need to re-run the data conversion and publish the Linked Data through their own site rather than DataIncubator. This makes adoption much, much easier.

Community Norms

Alongside addressing these procedural aspects of the data conversion process, the DataIncubator project also encourages a number of useful community norms that will hopefully improve the quality of the converted datasets.

The first of these is to ensure that there is a sufficient amount of both linking and attribution. Every dataset within the umbrella project should reference its original sources. This should not take place just at a high-level, such as within in the corresponding Void description: http://rdfs.org/ns/void/. Instead, references should be deeper so resources can be associated with, for example, the original web pages that describe them. This ensures that there is a clear path back to the original source of the data. Attribution—in various forms—is an important community norm in its own right, but it is especially important in the context of converting and re-publishing an existing dataset. We want to ensure that the original curators of the data don't think that the community is trying to appropriate or steal its work. Quite the opposite, we want them to embrace it.

Attribution—in various forms—is an important community norm in its own right, but it is especially important in the context of converting and re-publishing an existing dataset.

The other norm relates once again to sustainability. Links to the data should be stable, but how do we achieve this if the data will ultimately be removed from the DataIncubator site and moved to another domain? The proposal here is that as data is migrated to its permanent home, redirects will be put into place to ensure that web browsers and semantic

web agents can follow the links to their primary source. Every effort will be taken to ensure that links don't break.

What's In It?

The DataIncubator project already has a wide range of datasets available:

http://nasa.dataincubator.org – Data about satellite launches from 1957 to the present day

http://ol.dataincubator.org – An attempt at improving on the modelling of the OpenLibrary data

http://discogs.dataincubator.org – A conversion of the public domain Discogs music database

http://periodicals.dataincubator.org – Data on thousands of academic journals and publishers

http://airports.dataincubator.org – Facts and figures from the OurAirports.com website, cross-linked with Yahoo! Weather

http://jacs.dataincubator.org – SKOS conversion of the Joint Academic Coding System

There's a lot more that could yet be added to this list. My personal wishlist includes a conversion of the Prelinger Archives (http://www.archive.org/details/prelinger). This is hosted as part of the Internet Archive project and consists of over 2000 industrial, educational, travel, and propaganda videos published from 1903 to the 1970's. The content is completely within the public domain, so it's just begging to be converted. It would also be a great dataset on which to explore the modelling of media and media annotations in general.

Many gamers are typically very interested in statistics and data about the games they play. This is just one area of the Linked Data landscape that the DataIncubator project is hoping to help explore.

Currently, one domain with very little Linked Data is gaming, in all of its forms. For example there is a vast amount of community curated data about Lego, Lego sets, and Lego models. And what about all of the facts and figures that are routinely collected around online gaming? Data might be available through specific community websites, but what could be built if the data were more open, allowing the community to analyse and re-present this data in new ways?

It strikes me that games and gaming is an area that is ripe for exploration. There are many interesting dimensions to the data, and the communities are very engaged. Many gamers are typically very interested in statistics and data about the games they play. This is just one area of the Linked Data landscape that the DataIncubator project is hoping to help explore.

Leigh Dodds is Programme Manager for the Talis Platform

The Web of Data - Remixed!

By Michael Hausenblas and Richard Cyganiak of DERI


The Web of Data, and the LOD cloud in particular, has enabled access to a huge amount of online data. Now, with all this data at hand—what can we do with it? Certainly, one can use the data in applications, but quite often one wants to explore the data out there; somehow similar to what we now from the Web of Documents. Whereas with Google and co, we are used to dealing with documents; in the Web of Data we are concerned with entities such as people, books, genes, etc. These entities have properties (such as names, addresses, ISBN numbers, depictions) and relations to each other (author of book, works for company, and so forth). To explore the entities and their characteristics, we need new interaction paradigms, capable of addressing the distributed aspects of the Web data.

With Sigma (http://Sig.ma), we've created a visual Web Data aggregation and querying platform targeting entity visualisation and consolidation. Let's jump right into it: the user can start from a simple keyword query—e.g. "Stefan Decker"—to obtain a visual aggregate of the information about this entity. The result is depicted in Fig. 1:

> *With Sigma (http://Sig.ma) we've created a visual Web Data aggregation and querying platform targeting entity visualisation and consolidation.*

In the left column, the characteristics of the entity (a person, in our example query) are listed. On the right, the sources from which the data have been obtained are shown. Sigma uses Sindice, an index of the Web of Data, along with other indices from Okkam and Yahoo's BOSS. However, having a human in the loop is essential. To disambiguate the result, the user can remove certain sources (from the right pane) in order to narrow the result space, and, in case Sigma has missed certain facts, add sources on his own.

Essentially, Sig.ma can be used in three main ways:

- As a Web of Data browser to provide an integrated view: the user can start to form a query and then continue the browsing by following links to other resources.
- As a semantic "entity profile builder", potentially leveraging data from all over the Web. The resulting entity profile can be graphically customised and embedded into web pages. Entity profiles are always "live" meaning that changes in the original data sources will be reflected.
- Via its API for retrieving property value descriptions about entity or multiple entities. For example, querying for "affiliation@Stefan Decker" will return the said data in RDF, JSON, etc.

In the Figure 2. the last two features are demonstrated. It's creating a permalink and exporting certain characteristics of an entity in Sigma.

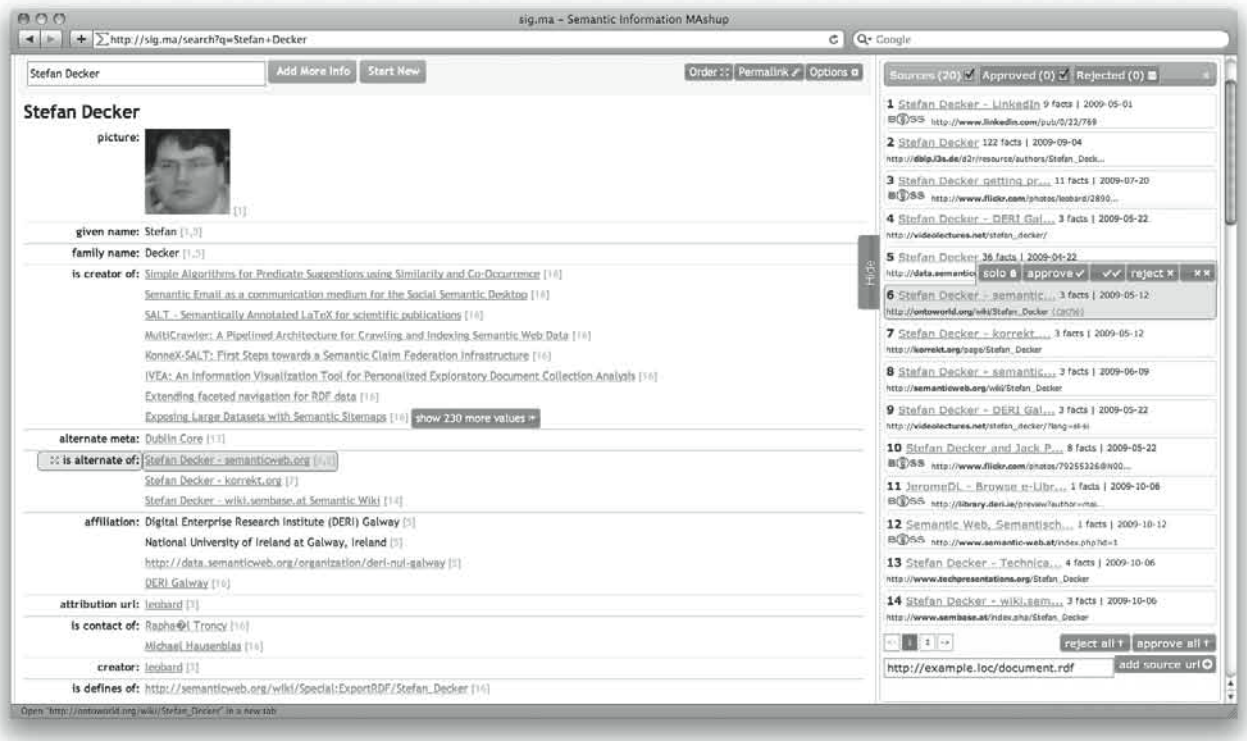


Fig. 1 Browsing the Web of Data with Sigma.

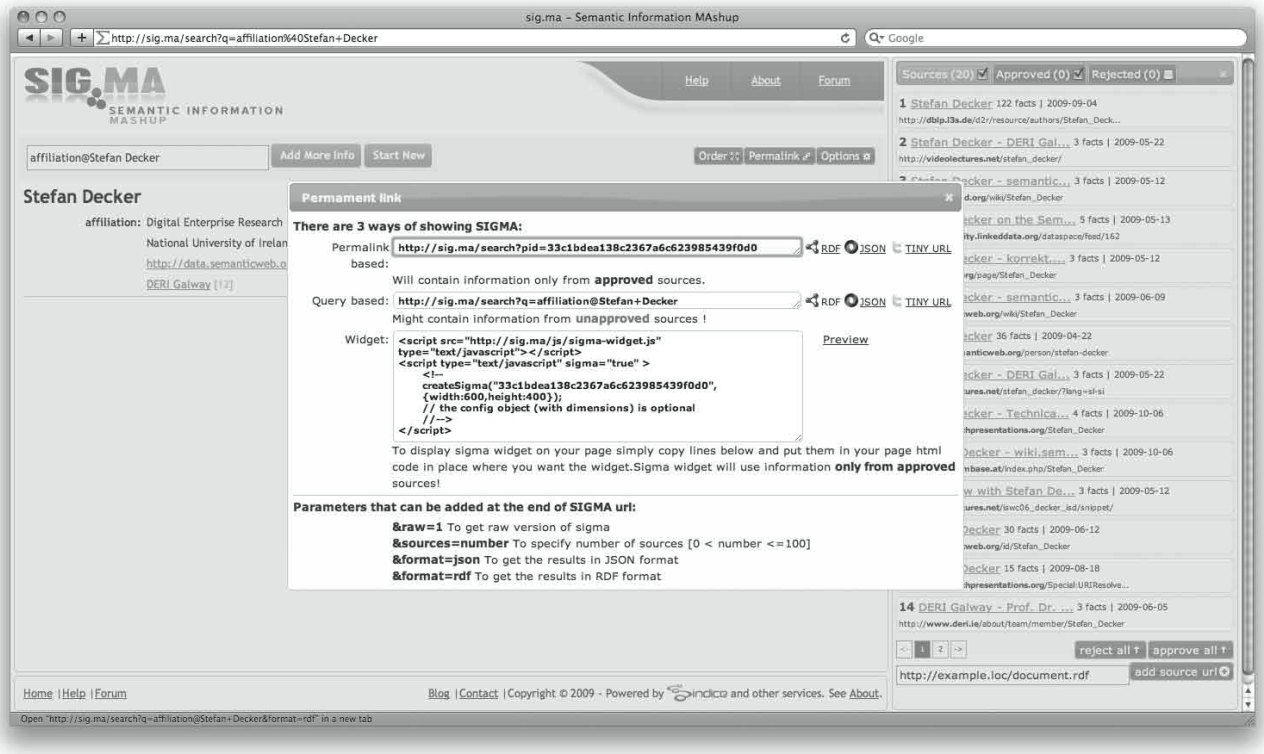


Fig. 2 Re-purposing Sigma's results.

To learn more about Sigma and its features, you might want to watch a screen-cast http://vimeo.com/5703809) where we explain in more detail how Sigma can help you harnessing the Web of Data.

Behind the Curtain

How does Sigma create the profile of an entity? First, the resource descriptions are merged into a single entity profile. What this means is that the key-value pairs from all resource descriptions are combined into a single description. Several techniques are employed, both empirical and following the ontology description as well as best practices and heuristics to extract property names. Next, we apply a manually-compiled list of approximately 50 preferred terms. For example, we replace the property names "work info homepage", "workplace homepage", "weblog", etc. with the preferred term "web page". After consolidation, the properties are ranked using the frequency of the property—the number of sources that mentions the property. This ends up pushing generic properties such as "label" and "type" to the top.

For key-value pairs where the value is not a literal value (such as a name or a date), but a reference to another resource (usually by URI), a best-effort attempt is made to retrieve a good label for the resource using methods such as fetching the remote RDF resource and looking for properties—such as "foaf:name", "dc:title" or "rdfs:label". It also treats the URI as a text and cleans it accordingly. Labels and values are retrieved and displayed incrementally, yielding a smooth and efficient interaction.

> *While we have already provided a valuable tool with Sigma, there are certainly a number of challenges remaining.*

What's Next?

While we have already provided a valuable tool with Sigma, there are certainly a number of challenges remaining. On the one hand, we have to deal with noisy data stemming from the live heterogenous data sources we are using. On the other, spam—as recently pointed out by Ian Davis (http://iandavis.com/blog/2009/09/linked-data-spam-vectors)—can lead to a potential trust problem. For now, by listing the data sources explicitly, we can rely on the user to assess the trustworthiness of data sources contributing to an entities rendering. However, this area likely requires further research.

Richard Cyganiak is a Semantic Web evangelist and a researcher and engineer at DERI.

Michael Hausenblas is a Web of Data researcher, coordinating DERI's 'Linked Data Research Centre' (http://linkeddata.deri.ie/).

Hungry, Passionate?

Apply yourself at Talis

Talis is a dynamic, innovative, established technology provider. The success of our company depends on the individuals that we employ. So, if you are adaptable, like a straightforward and plain speaking culture, have the stamina to maintain high standards over long periods of time, and care deeply about what you do, Talis could be the place for you.

CREATIVE SOFTWARE DEVELOPER
We are currently looking for a creative software developer who will add value to our team, our products, and our company. We want someone who isn't satisfied with just solving problems; he or she will want to bring creative and innovative ideas to the table - and turn them into reality. Our type of developer will be proficient in using PHP and comfortable using other scripting and programming languages. Salary £35 - £45k dependant on experience.

WEB APPLICATION TECHNICAL LEAD
We have an opening within the Library Division for a Web Application Technical Lead to drive the continuing development of Talis Prism, our next generation library catalogue product. This role will not only make technical decisions around the architecture, design and implementation of the product, but to be successful you will need to possess a high level of passion and commitment in what you do. Salary £37 - £47k dependant on experience.

APPLICATION DEVELOPER
We are currently looking for an Application Developer who is not only able to solve problems but will own what they're doing and take it to the next level. You will need to have the confidence to switch between technologies and possess a range of programming languages – we are not looking for those who are focused on a particular one. We want someone who has the ability to build great software, coupled with the willingness and drive to continue learning. Salary £35 - £45k dependant on experience.

Talis provides a series of very competitive benefits including a contributory pension scheme, healthcare cover, optical and dental cover, a childcare scheme, all within a highly flexible working arrangement. We envisage you working on average a 36.25 hour, five day week. Initial annual leave entitlement is 22 days and this rises by 1 day a year up to 27 days. If you are interested in any of the above roles, please email careers@talis.com with your CV, covering letter, and salary expectation.

Talis, Knights Court, Solihull Parkway, Birmingham Business Park. B37 7YB, UK. www.talis.com, email: careers@talis.com.

shared innovation™

talis®

Please visit www.talis.com/careers for more details. If you want to have fun at work and enjoy being highly visible and accountable for what you do, we would really like to hear from you. Based in the Midlands, we offer an attractive range of benefits including home and flexible working to support your whole life balance. Check out some of our employee blogs on www.talisians.com.

Are Clouds Green?

By Paul Miller, Cloud of Data

On the face of it, the answer to this question should presumably be a resounding: "Yes!" With far more efficient utilisation of available compute resources, economies of scale in staffing and infrastructure, and a tendency to site the bigger data centres in areas blessed with natural cooling and abundant green-ish power, the answer is surely obvious. How could a Cloud be any less Green than all those on-premise data centres, full of woefully under-used computers that endlessly pump waste heat toward a roof filled to overflowing with thermal exchangers and power-guzzling air conditioning units?

Surely, as world leaders fall over themselves to exhort someone else to make the painful decisions ahead of the summit in Copenhagen, Cloud Computing is the technology community's contribution to making sure the world is still here for our children to inherit and enjoy?

Chris Thorman has an interesting post that is unusual in trying to apply some real numbers to the discussion. In it, he compares the estimated direct energy costs of running a particular medical application on-premise to the same application being hosted by a third party.

The conclusion, at least based upon the factors Chris chose to measure, is unsurprising;

"93% reduction in overall energy consumption… using SaaS EMR software over on-premise software!"

Simon Wardley, one of the Cloud's more effusive cheerleaders, offers a less straightforward analysis in a recent video conversation with GreenMonk's Tom Raftery.

Simon agrees with my opening assertions that compute resources in the Cloud should be more efficiently utilised and therefore cheaper to run on a per-unit basis. However, Simon goes on to suggest that the near-ubiquitous availability of computation will lead to a profligate rise in usage. We might, as Chris Thorman suggested, complete a specific task more cheaply and efficiently than before, but we'll run more tasks without giving sufficient thought to the consequences. Just as lights and gadgets get left on all over the house because the power is just there (and too cheap to worry about in most cases), so too will 'available' computers be squandered. It will be easy to call upon a computer when required, and application developers will be quick to spot the opportunities to put those spare cycles to work in completing myriad small, individually inconsequential yet collectively shockingly wasteful tasks. We see this with desktop computers today, casually devoting once-scarce resources to the generation of 'eye candy.' On an individual basis the waste is small and perhaps easily ignored. Scaled to the level of the Cloud it will surely become far more noticeable, far more measurable, and far less easy to justify.

Is this the Cloud's 'fault?' No, of course not. Instead it appears baked into the attitudes with which we view commodities. So whilst the technologies behind the Cloud might make Green efficiencies possible, human nature may very well pull in exactly the opposite direction. Which big Cloud provider will be brave enough to implement a 'Green tax,' charging customers a premium for 'wasteful' computing? And who will decide what's wasteful, anyway? Even the Cloud providers don't always help themselves. Headline figures from 'green' data centres such as the behemoth recently opened by Microsoft in Dublin are certainly impressive, but it's effectively impossible to discover the consumption or efficiency of individual machines or units of computation. Is the secrecy simply about competitive advantage, or does this new generation of data centres have something to hide?

On an individual basis the waste is small and perhaps easily ignored. Scaled to the level of the Cloud it will surely become far more noticeable, far more measurable, and far less easy to justify.

And Simon's final answer to the question, 'is Cloud Computing Green or not?,

"Yes and No."

Indeed.

Customers of Cloud and traditional hosting company Rackspace would appear inclined to agree. When directly asked as part of Rackspace's latest Green Survey,

"Do you view cloud computing as a greener alternative to traditional computing infrastructures?"

- 21% agreed, "Yes, cloud computing is a much greener alternative"
- 35% were "not convinced of the green benefits of cloud computing"
- 25% reckoned that there was "too much hype around the green benefits of cloud computing"
- 19% suggested that "the true green benefits of cloud computing have not yet been realised."

I'm actually surprised that the pro-Green numbers weren't higher. Did respondents not understand the widely used efficiency argument, did they not believe it, or had they moved past it to embrace Simon's arguments about our penchant to waste anything abundant?

There are some fascinating figures in the survey, although like many similar exercises I'm left with as many (different) questions as when I started.

So… Are Clouds green?

I think I'll follow Simon in hedging with 'Yes and No.'

What do you think?

Paul Miller is founder of cloud computing and semantic technology consultancy, the Cloud of Data. http://cloudofdata.com/

Chris Thorman's blog post: http://www.softwareadvice.com/articles/medical/saas-v-on-premises-which-one-is-more-green-1092209/

Tom Raftery's video interview with Simon Wardley: http://www.youtube.com/watch?v=zzAEBokyyBs

Rackspace Green Survey: http://www.rackspace.com/downloads/surveys/GreenSurvey2009.pdf

Juice: A partner for Linked Data

By Richard Wallis, Technology Evangelist at Talis

Sir Tim Berners-Lee has been encouraging us to get our data out there.[1] The British government are more than following his simple advice[2] by publishing some of their data through SPARQL endpoints, with help from the Talis Platform.[3] The BBC has been sprinkling RDFa throughout their web pages and Wikipedia info panels are being published as RDF via DBpedia. Early days yet, but Linked Data tendrils are starting to creep across the web like virtual ground-elder across a springtime flower bed.

Unlike those weeds, however, these Linked Data tendrils are starting to be used to deliver beneficial results. Building upon already published data, from reliable sources, a different breed of application is starting to emerge. These applications are making use of light-weight and simple Internet protocols to query stores of data—described and curated through community-formulated ontologies. While these applications grow, a different way of thinking about application development is also beginning to emerge—slowly eroding the, "pull as much data in to my application as practicable, to provide the best user experience" approach to design. The more open-world view of drawing resources from wherever they are best created and curated, to add as much value to the user experience as possible is in the ascendence.

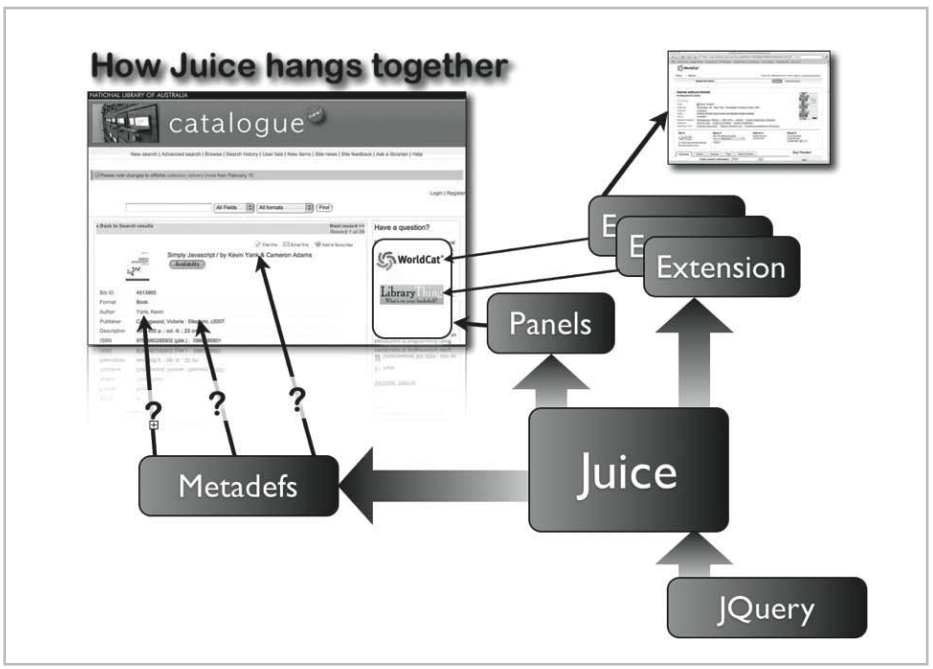
> *The BBC has been sprinkling RDFa throughout their web pages and Wikipedia info panels are being published as RDF via DBpedia.*

App Garden

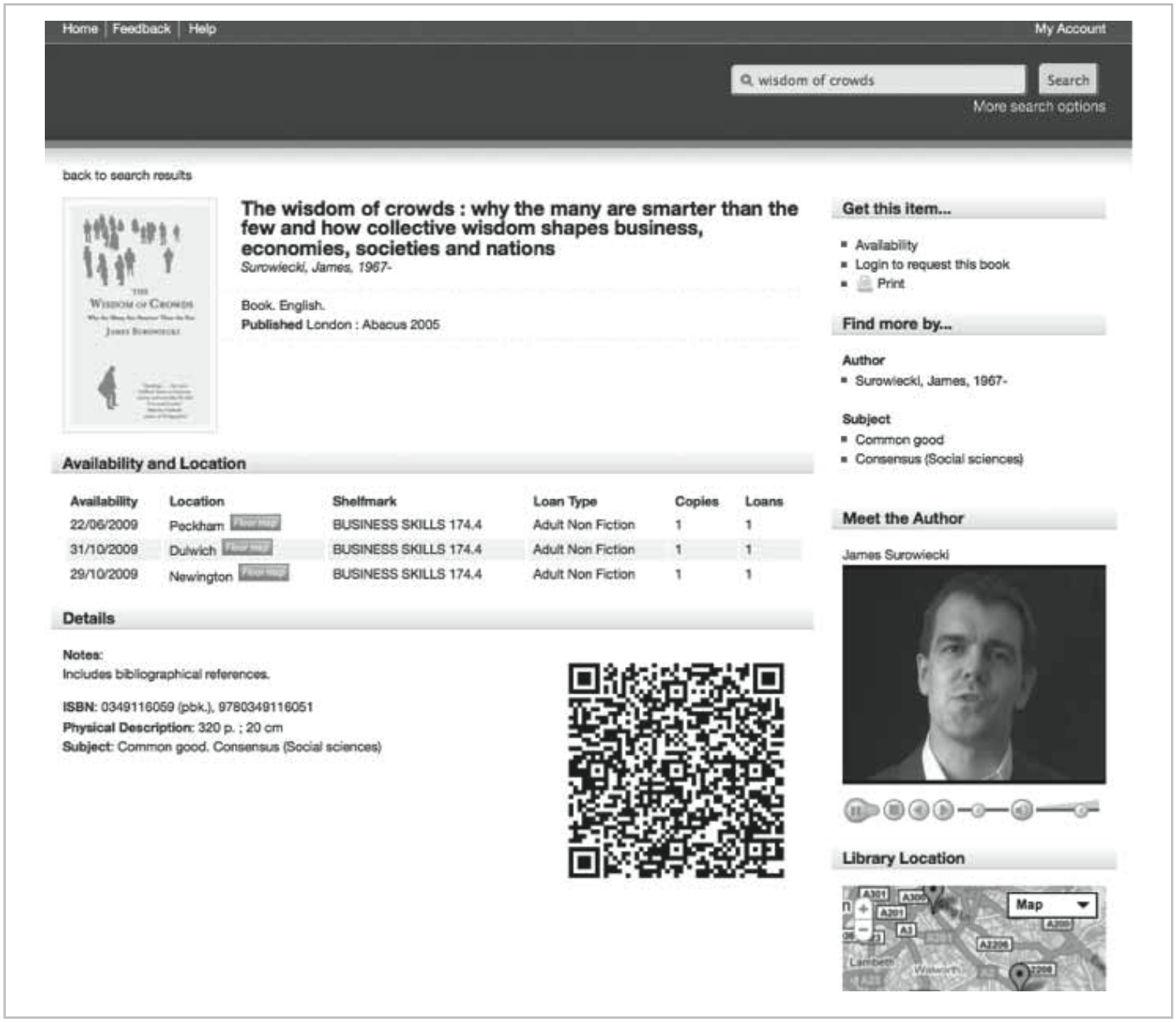
The National Library of Sweden[4], like many national libraries, uses an internationally-agreed set of subject headings to help catalogue their books. The Library of Congress in Washington DC have maintained and published these for many years, and the Library of Congress Subject Headings (LCSH) were an early example of a dataset exposed using Linked Data principles. This has enabled the Swedish Library to be one of the first to link out to this service instead of copying them in to their system. Although a simple example, you only have to extrapolate it out to all the other libraries in the world to imagine the consistency and efficiency benefits that will eventually emerge.

It is one thing to pull in useful data from elsewhere on the web; it is yet another to consciously delegate out the management of data that is core to your application to a service operated by someone else. Yet this is exactly what the developers of the new BBC Wildlife Finder[5] site have done. As well as partnering with several organisations to provide detailed animal distribution, conservation, adaption, and behavioural information to support their stunning video and audio content, they are using Wikipedia to provide general background information. When they identify gaps in this information they will be contributing to Wikipedia to fill them, instead if embarking upon an internal content management exercise. Not only will they be increasing the quality of the BBC site but increasing the spread and quality of Wikipedia for the benefit of all.

Linked Data and the principles behind it, the openly published datasets, and the change in philosophy towards application development all bode well for the development of the future Web and the applications that will deliver it to the wider population. Great news for the application developers, but what if you are not an application developer? What if you just manage an application? What if you are one of the thousands of people who keep sites and services provided by third parties running for their organisations? What if all you can do with your application is add in a few lines of html? What if you are not confident enough to do any more than that? What if your application supplier does not allow you to do anymore more than that? How do these folks take advantage of the obvious benefits of the Linked Data web? This is not a new problem. There have been many a site or system manager who has looked longingly at Web 2.0 extensions on other sites. A Google map of locations, a Twitter



1. http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html
2. http://data.hmg.gov.uk/
3. http://www.talis.com/platform
4. http://libris.kb.se/
5. http://www.bbc.co.uk/wildlifefinder/



feed, embedded ratings and reviews, how do I get one of those on my site!

No Green Fingers?

This is the background that provided the drive for the launch of the open source Juice Project. Juice—Javascript User Interface Componentised Extensions—is a project with the mission to make it easy to share user interface enhancements and extensions. The germ of the idea was formed following the launch of API access to the Google Books service. Within a few days of it's launch, there were a handful of sites providing links back to Google so that you can view the scanned image of a book that is being referenced. A few weeks later saw a few more sites, but not many. After a few months, some had taken advantage of the ability to embed the book viewer in your page, removing the need for the user to visit the Google site. Nevertheless, in total there were surprisingly few sites taking advantage of this powerful extension in their offering. Although the API is simple to understand, each site had to understand how to embed it in to their user interface. Innovation was effectively being stifled because of the lack of consistency between interfaces.

The proposition behind Juice is that, if everyone can use the same simple steps to embed such information, the whole community would begin to benefit from the enriched content—perhaps even encouraging further innovation. Juice does this by hiding the complexity of interfacing with a user interface or an external service behind simple copy & paste steps.

To introduce Juice in to your interface, you first need to paste three JavaScript calls in to your html:

```
<script type="text/
javascript" src="http://
example.com/jquery.js"></
script>
<script type="text/
javascript" src="http://
example.com/juice.js"></
script>
```

Two of these calls pull in powerful JavaScript libraries, the third is your a site-specific configuration file which will initially be empty. At this point, a—possibly nervous—site owner can simply check that introducing Juice has not damaged the operation and performance of the site they look after. Again, copy & paste comes in to play when placing extensions in to the site-specific config file: one line each to load the individual extension JavaScript file and then to call it.

```
juice.loadJs("http://
example.com/extendedbyJuice.
js");
new extendedbyJuice(juice);
```

This pattern then continues for all other extensions. (Some extensions will obviously need some controlling values passed to them: for example, Google Books needs an ISBN.) This information is pulled from the page content, also in a shared way. The first developer to add Juice to a particular interface type will create a definition script identifying where information is held in the page html. By including a call to this, a site owner can reuse that work and enable any of the extensions that require it.

Community Gardening

The community forming around this new project generally consists of three categories:

Those JavaScript and API experts others wish to emulate—creating the core Juice code and extensions.

Those who are comfortable writing simple JavaScript and html—creating new extensions and interface definitions.

And the remaining, vast majority who want simply to share in their innovation.

This brings me back around to the way that the massive benefits of Linked Data and Linked Data principles can be brought to the vast majority of the population who do not know the difference between a web browser and the Internet, let alone what Linked Data is. If every Linked Data / Semantic Web enthusiast took a leaf out of BBC Nature's book and surfaced this type of value in their service, we would still only be scratching the surface of the web being consumed by the wider population. Most web sites are about searching for things and only displaying internally-held information about those things—effectively dead-ends in the discovery process. The BBC Wildlife site is an excellent demonstration how those dead-ends can be transformed into jumping-off points in to exploring related information and things, both on their site and the wider web. Unless you are starting from scratch and building a new site, this sort of innovation is difficult to emulate. The Juice Project could have an important role to play here by enabling current sites to turn their dead-ends into their own jumping-off points without a major rewrite of the code behind them.

Some Linked Data evangelists are encouraging us to embed RDFa in to our web sites. Juice is ideally placed to capture that data, or simply capture some displayed html, and use it to power in-browser extensions that will either embed resources such as video and audio into the page or start the user on an exploratory journey across the Linked Data web.

The Juice Project can be found at juice-project.org. Join in and help spread the data web to where users can see and use it.

Richard Wallis is Technology Evangelist at Talis and the founder of the Open Source Juice Project.

Tom Steinberg Talks with Talis about the Work of MySociety

Podcast with Paul Miller

Paul Miller: Hello, and welcome to this podcast with your host, Paul Miller. Today I talk with Tom Steinberg of MySociety. We discuss the work of this charity and some of the ways they have been agitating for a number of years for greater transparency with respect to government and government data.

We look at some of MySociety's projects and consider the next steps as government data becomes increasingly available for all of us.

For more information on this show and others in the series, please see the show notes on our website. Thank you.

Tom, thank you very much for joining me for this podcast today. Could you tell our listeners a little bit about yourself and your own background, please?

We look at some of MySociety's projects and consider the next steps as government data becomes increasingly available for all of us.

Tom Steinberg: Well, I'm the director of MySociety, which is a non-profit based out of a charity in Britain. But my background: I'm a policy analyst and a one-time systems administrator, so I have a foot in both camps.

Paul: That's an odd combination, the systems admin side and the policy. Most policy people you meet wouldn't know one end of a computer from the other.

Tom: I think I would agree with that. In some ways it was a great bit of good fortune in my early career, really. I could have been into policy and surfboarding, and that wouldn't have been quite as useful.

Paul: I don't know. It depends where you are, I suppose. Possibly not that useful in London. MySociety, what's that trying to achieve?

Tom: It's trying to achieve two different things. The first and higher priority is to use the Internet to build websites that give citizens simple but very tangible benefits in the civic and democratic parts of their lives - simple benefits and tangible benefits being things like finding out who my politician is, finding out how I can write to them, finding out how I can get a pothole fixed in my street, finding out how I can get some information out of the government that is otherwise not public. That is mission one, serving the public, and that's the charitable part of what we do primarily.

The secondary part - which I suppose I would argue is also charitable, but it's different - is persuading or teaching the public sector and the voluntary sector how they can themselves use technology best to deliver these sort of crunchy, tangible benefits rather than brochures or websites and initiatives that hit all the buzzwords but don't do anything.

Paul: You've done a lot of both those missions through a series of project websites which have been running for some time, a number of them. What kind of things have those project sites tended to focus on?

Tom: Our biggest website is TheyWorkForYou.com, and that focuses on making it easier to understand what's going on inside Parliament if you're not a parliamentary insider. It's not a political blog: it doesn't write the tittle-tattle. It's just a whole lot of data that has been presented in a mostly automated fashion so that it's very easy to understand, you know, "My MP voted mainly against foundation hospitals," in that sort of very simple plain English.

But it's very neutral in its outlook, and it has an enormous amount of information in it because it has everything every MP has said back to 1935, lots and lots of votes, lots of other things. So that is an example of the sort of service that we have built.

Our biggest website is TheyWorkForYou.com, and that focuses on making it easier to understand what's going on inside Parliament if you're not a parliamentary insider.

Paul: How easy has it been over the years to access the data that these services tend to rely upon?

Tom: Well, it varies really. The main story has been not very good. When the original volunteers who built TheyWorkForYou took Hansard, the record of Parliament, they took it off the parliamentary website and started making it more easy to use, and search, and permalink in some... The first thing that happened was that Parliament started the procedures for suing them, and they then changed the policy on that and stopped. But nevertheless, that I would suppose is an indicative approach to what it means to be an innovator in this field and the general historic reaction that government has given to you.

Now things are a little bit different because we have some champions within government who will go and fight for your side, but those are still relatively few people within a relatively big machine.

We have some champions within government who will go and fight for your side, but those are still relatively few people within a relatively big machine.

Paul: Indeed, indeed. When MySociety started out, it was something of a lone voice and has since been joined by other reasonably high-profile activities like, for example, "The Guardian" newspaper with their "Free Our Data" campaign. We're beginning to see moves on both sides of the Atlantic and elsewhere in the world to make sure that large amounts of public data are more easily available. We've seen, for example, Data.gov in the United States, the work of the Sunlight Foundation, and we've got similar moves apparently afoot here in the UK.

Do you think we're actually going to see lots of unencumbered public data, or will the government machine find ways to make it less useful than it might otherwise be?

Tom: Well, my main theory is that we will see lots of it that nobody's interested in. I was involved in writing the Power of Information Review, which was the government here's first attempt to take the issue of public data in the Internet age more seriously. It struck me halfway

through this process that there are two quite different philosophies of liberating government data that you can have. There is essentially the one that perhaps sounds better in terms of the headlines where you say all shall be open, and you pass out some sort of mandate.

The other is one where you say, "We will work as hard as we can to open anything that you ask for." Or even to be tougher, "All shall be open that the public asks for," which is different from "all shall be open."

I'm very much in favour of the secondary approach. I slightly fear that broader mandates that just say all departments shall liberate everything will lose some of the focus because there will be a lot of sound and energy.

ask for something on here, we will go and do whatever we can, or we will enact some new legal power that doesn't exist at the moment to go and extract that, or we will even just have some discretionary funding that we can spend on getting that kind of information out there.

That is one really big gap. The other big gap is that this site actually simply needs to be publicized where there is data. The web people run up against brick walls and find they can't get the data they want isn't on the Office of Public Sector website. It's on things like rummaging around the insides of the Transport for London website looking for some data.

Now from personal experience, if you try to get everything out of Transport for London you

So I'm very much with Tim Burners-Lee on this. Go, get what you can, leave a note saying "please improve this." But don't, for God's sake, hold up the process while you republish your data in the right way.

It will look like there's been a great deal of success just because 10,000 data sets get put on some server, but they might be 10,000 data sets that were already available everywhere else, 10,000 data sets that were not the ones that the public is screaming for.

So in my mind, the right way of solving this problem is still to say—the ultimate policy goal that no one in the world's promised yet, what I'd like someone in the world to promise, would be: if you make a case to us that there is something of economic or social value that you can't do, we will bend over backwards to make sure that it gets to you for the strength of our country.

That therefore is my take. It's not a bad thing that you might have a blanket "all shall be open" mandate at all. The freedom of information mandate under the Freedom of Information Act is that kind of mandate.

But I think in a time where there's not very much money knocking around for an issue which is in some ways less emotive than freedom of information, keeping a bit of your powder dry and targeting it where people really want it is my chosen solution.

Paul: How would we set about deciding what data was wanted?

Tom: One thing you can do is simply ask people, and the Office of Public Sector Information have a data-unlocking tool on their site, which is a place where you can essentially petition for certain data sets. You can put your name against other people's requests. That is only a very beginning though, because it lacks two things. One, it lacks a really strong stick. It lacks something that says if five or ten people will be sorely disappointed. And a mandate from central government that says if you publish data, essentially, someone on your site—that has to be a really clear link that says if you want something from this site that isn't available, go to the Office of Public Sector Information, file a request, and we will then follow up with hopefully strong statutory powers. That's the right way of doing things.

So you need to use resource, and you also need some publicity that the government runs this process in the first place.

So you need to use resource, and you also need some publicity that the government runs this process in the first place.

Paul: OK, so a lot of public-sector bodies here in the UK currently have a designated contact for an FOI, Freedom of Information request. You're suggesting then some similar mechanism being in place for data requests.

Tom: Well, no actually, not quite, because that means hiring. There's a hundred thousand public bodies in Britain, and that means hiring a vast number of new people. I actually just literally mean a prominent link that is the same on all government sites that says, "If there is some information that you need from this site that you cannot extract, perhaps because it's against this organization's policy, go here, file the record, and we will go and chase it," basically.

Paul: Right, OK. So very much putting the onus on chasing back on to OPSI then.

Tom: Yeah, who would of course correspondingly need greater resources for it. But you know, I've met public bodies where I've had a whole bunch of meetings to get one dataset, been fobbed off, and literally just never got anywhere. And so just passing that on to someone whose job it is doesn't mean anything at all, because their job very well could be interpreted as not giving you the data, and that doesn't get us anywhere.

Paul: Indeed. So this data being made available, are there any requirements around how the data is made available when someone has decided to let you have it? Or will anything do?

Tom: I'm a pragmatist really, and I would rather anything, than something that's in six month's time and is better, or in a year's time and is better. Clearly there are emerging standards. The right way of doing this is evolving pretty quickly, and is pretty straightforward. But it's pretty unreasonable to think that a lot of public-sector bodies will do things like that. So I'm very much with Tim Burners-Lee on this. Go, get what you can, leave a note saying "please improve this." But don't, for God's sake, hold up the process while you republish your data in the right way.

Paul: OK, so what happens when all of this data becomes available? Does the world change? Does our engagement with government change fundamentally, or do things simply get incrementally a little bit better?

Tom: Well, I think we can expect a mixture of a big load of not much change, and enormous amount of virtually no change, and one or two quite sort of dissonant and surprising things. That's my expectation from the history of the Internet about how datasets get used. I know that, for example, there's some criticism coming from Gartner about the various US experiments in doing mash-ups basically—match-ups in government and public data. And I think this is a bit hard, but in some ways they're reasonable in saying: "Well, look. You said we were going to open this data, and the world was going to change. But we don't even have anything that awesome. Like, there's nothing out here that any average American would want to go and use, off-hand."

I would expect that to happen very rarely. I think that criticism is rather naive, really. I would expect that to happen one time in a thousand.

But probably the reason it's worth doing is because that one time in a thousand will be amazing. And I still think in some ways perhaps, one of the preeminent examples here is Fix My Street. Everyone understands what Fix My Street is. It's a website where you stick a pin in a map, and it sends your problems to councilmen and helps get it fixed, and that's based on various government datasets.

But there are not actually enough examples out there now about things that really use data in a way that really, really matters to non-techie people. And I don't think it's fair to expect there will be a whole load in the short run. But nevertheless, there will be none without this liberalisation process. And I'm pretty sure someone will pick up one or two datasets and really run with them enormously. And that is almost certainly why this is worth doing.

I'm pretty sure someone will pick up one or two datasets and really run with them enormously. And that is almost certainly why this is worth doing.

Paul: And it almost may be that a lot of the improvements are actually largely invisible. They're about increasing flows of information between national government and local government, perhaps. Or interactions between service providers and government departments. It may be that the public never actually sees a lot of what going on.

Tom: Yeah, so one thing I would know if I was a minister in a new government, was that I'd be terrified about all the things going on in my department that I didn't know about. And I would also disbelieve anybody who said to me, "Oh, all you have to do is ask your civil servants, and they'll go and find out what's going on and tell you." I would think that it's not possible that they could possibly know everything that's going on. Nice kind of data flows that are exposed to the outside world so that people build things like they work for you with them, and then user bases that trawl through this. That seems to me a much better way to a minister finding out what's going on under the remit.

I mean, we see this in relation to They Work For You. We saw that hundreds of people every week who are inside Parliament looking at They Work For You. Why aren't they just looking at the official record to find out what goes on in Parliament? The answer is having released that data and us having processed it, it's much better than anything they've got, in that same way.

So I think one of the very strongest arguments, if I was a minister and I was into this agenda, and I thought it was right for the public, the taxpayer, and I thought it was right in terms of creating innovation, but I think it's more right in terms of creating services that would save my backside from a massive scandal I had no idea was coming later on, by opening this stuff so people could start processing while it was still basically the last person's fault, and when I could kind of say it wasn't me.

I think it will be very interesting to see how many ministers understand the risks to themselves of keeping their data closed, because they're not going to be the ones that benefit from it.

Paul: That's an interesting point. So as well as protecting your reputation or pinning the blame on someone else, does this increased transparency save us money or cost us money, do you think?

Tom: Well, at the moment it's almost certainly a hypothetical question. Clearly there's some direct costs and there can be some direct savings if you expose things that are clearly waste. The government itself, I believe, has long since identified 10 or 15 billion quid of what it thinks are essentially waste savings out there. And certainly the opposition and the hostile press probably think it's a lot more. One of the longstanding questions in government is, how on earth do you spot this, given that it's so big and you spend so much money? My recommend would be if this sort of transparency helps you even shave 1% or 2% off a waste bill of a billion quid, then you have almost certainly paid for every initiative you could possibly do in this field about ten times over.

So I think I would be hugely surprised if you were to do a study in five or ten years and it turned out that this didn't lead to net savings.

So I think I would be hugely surprised if you were to do a study in five or ten years and it turned out that this didn't lead to net savings.

Paul: I guess we'll find out. It'll be interesting to come back and revisit this in five years time, and see. I think I agree with you. One of the interesting projects you did recently at MySociety was Mapumental, which I think appealed to lots of people I showed it to, partly because it's so visual. It's immediately apparent what's going on. And this project is presumably where those comments about getting data from Transport for London came from. Is it?

Tom: Mm-hmm. Yep, yep. Ultimately we got the data from the National Public Transport Repository. And what was disappointing about the discussion with TFL, but worth mentioning in this interview, was not that they weren't nice, it was that the people sort of in charge of the kind of data thing knew so little about their own data structures. They were so substantially outsourced below them, that they actually denied it was technically possible to produce a copy of the time tables, which left me slightly agog. Clearly on a time table website you have data for time tables.

You might expect your barrier would be refusal to play a game with the government because of the "not invented here" syndrome. Or might expect people would say it's too expensive.

And so that was an interesting sort of example of one of the barriers you might expect. You might expect your barrier would be refusal to play a game with the government because of the "not invented here" syndrome. Or might expect people would say it's too expensive.

Finding that people literally understand the issue of public data so minimally, they don't even understand that they are in a position to contribute when someone is sitting on the other side of the table saying, "You have this stuff. It would be really good if we could have it." And then literally going, "Uh, I don't think we can actually do that. Like, I just think it's not possible."

That is a new barrier, a different and strange barrier, one that would be very alien to most kind of technical people who are listening to this. And yet it's sort of indicative about the fact that you can't plunge into this field assuming that you know everything about why data isn't made public.

Paul: What do we do about that?

Tom: Well, I tend to think that a stronger OPSI with statutory powers to sort of almost go and cease, and maybe some money to make the seizure process less painful for those at the other end of it. I think that's almost certainly the right way to go.

Paul: OK. It would be interesting to see, again, whether government is prepared to make that kind of change. So I mentioned Mapumental

there, and I'll include a link to the video and things so people can see it. But, what was it trying to do as a project?

Tom: Well, it is like everything MySociety does. It's about a simple, tangible benefit. And the simple benefit here is, where should I live, or where should I work, or where should I send my child to school --some of the biggest decisions we have to make in our lives. Which are made complicated by the fact that, especially in cities in Britain, transport system is extremely complex. It can be very difficult to know where you can live and still have acceptable commute to any of these places. And that's the main goal, really. There is a second aspect—kind of proof of concept technology showing that these things are possible. And also an aspect of driving the repetitive message home in relation to public data, which is that public data is fabulously valuable. It's being abysmally wasted in some of the ways it's being used at the moment. And it's really, really important that it gets out more easily.

It's about a simple, tangible benefit. And the simple benefit here is, where should I live, or where should I work, or where should I send my child to school --some of the biggest decisions we have to make in our lives.

That data costs us the best part of 10,000 pounds to get a hold of, and it's timetables. It's manifestly public interest, and yet we as a charity had cough up a lot of cash.

Paul: And it also includes house price data as well, doesn't it?

Tom: Which you also have to pay for, so...

Paul: Yes. And it's just London at the moment. Are there any plans...?

Tom: No, it's the whole of the country.

Paul: Is it? All right, sorry. Watching the video, that was just London, wasn't it?

Tom: Mm-hmm.

Paul: Yes, OK, good. So we've talked quite a bit about increasing access, increasing transparency. Do you think, honestly, that the government or the opposition understand the opportunity here?

Everyone has to be conservative about any government that comes in next in a time like this, because no one will have any money. But nevertheless, I suppose I am cautiously optimistic.

Tom: I think that there have been a handful of people in the present government who understand it, but they are too small and the government is in too much of a late, late era sort of death spiral to be really very serious about reform now. It's just an ugly honest --it's a big open secret. The nice thing about not being politically partisan, by the way, is I can just say that as a fact that everybody knows rather than a matter of kind of opinion. The opposition, who I am also kind of non-partisan about, appear to be talking with more energy. They want to copy everything that Obama's done that they probably can for obvious reasons, and Obama talked an awful lot about this. And so they're talking about it, too.

Everyone has to be conservative about any government that comes in next in a time like this, because no one will have any money. But nevertheless, I suppose I am cautiously optimistic. And perhaps that's no surprising, because nearly any party that's been in opposition at this point in the 21st century has been relying less on the comforting bureaucratic structures of the Whitehall office. Your driver, and your private office, and your red box, and your beautifully typed memos done by a staff of people you don't see.

Whereas, if your in opposition, no matter where you are from in the political spectrum, you're going to have to have fended for yourself a bit more. You probably booked your own tickets on Easy Jet and done all that sort of stuff. And I think there is almost an unfair advantage in terms of getting Internet from the people who weren't shielded from it.

It's worth remembering that Tony Blair basically never sent an email during his time as prime minister. And can you imagine how genuinely difficult that would be for me or for you to do? But the civil service is really one of the few places left in Britain where you can really conduct your job reasonably well without having to turn on a computer.

And so that means the whole generational attitude difference is really different from anyone out there who hasn't had that sort of protection. That includes a lot of the opposition. So I think it's almost a generational thing than a partisan one.

Paul: And you think that the current opposition and the comments they are making, as you said, about copying some of Obama's ideas around Data.gov and its equivalent policies. Do you think if they get into government next year that they'll follow through on that? Or will they be sort of distracted by the economy that everything else will take a backseat?

Tom: Well, now you're asking me to get into political punditry. And I said before that clearly anything that costs any money is going to be very tough for anyone who wins next year's election, let alone the opposition, who we all expect to. I think asserting whether or not I have trust in a political party would be beyond what my job's worth as someone who essentially runs a charity. So, no comment.

Paul: Indeed, OK. So, MySociety started demonstrating what a possible, agitating to get access to some of the data. If all of that's happening, and it's becoming easy for anyone to pluck government data and do what they want with it one day, is there still a role for MySociety? Or is it time to move on to the next challenge in terms of transparency and data?

Our challenge is to always find, and our goal is to always look at what's newest, and what's next, and what nobody has built yet.

Tom: Well, I think that we are moving on all the time. We are not at all the same organisation we were a few years ago. The other thing is I think we are unbelievably close to the start of this transformation. And so the idea that we could say, "Well, that's Parliament and government fixed. On to something else," would be so laughably and self-evidently untrue that it would deserve being taken seriously. Our challenge is to always find, and our goal is to always look at what's newest, and what's next, and what nobody has built yet. We've never been into the game of building another site in a field that someone else already occupied. That relentless desire to be first is still strong. And given that there's so much stuff left to do in which one can be first in this field, unlike for example online shopping, that I just feel like we have endless amounts left to do.

I just noticed the call for proposals that we ran has got more than 200 ideas from the outside. So there's clearly plenty of people with plenty of ideas that haven't been built yet.

Paul: This is the call for projects from MySociety to tackle in the coming year?

Tom: That's right. It's just closed.

Paul: OK. Any thoughts at the moment in the sort of trends you were seeing in that? Or is it too early to say?

Tom: I haven't even begun to vet through it. It's a huge pile. Possibly the only thing is the number of people interested in government spending.

Paul: [laughs] I thought that was all done, wasn't it?

Tom: Spending? No, in this country there's almost no spending websites at all. In the US, you've got FedSpending.org and the government is moving on with things like Recovery.org and USAspending. Whereas in Britain we lead in some of these fields, compared with the States, in government spending we're really no where at all.

Paul: OK. So that may be the next big thing

Tom: You heard it here first.

Paul: You heard it here first.

Tom: Right.

Paul: So, thank you very much then, for that. It's been sort of a broad discussion of some of the things you've been doing, and some of your thoughts on it. Before we wrap up, have you got any closing thoughts to share? Any questions I really should of asked you?

Tom: I think not. I think you've extracted plentiful things from me, and I hope they keep your listeners entertained.

Paul: OK, good. Well, thank you very much, Tom. And I'll look forward to seeing what the 2009 projects turn out to be. Thank you.

Tom: Thank you.

The original podcasted conversation can be found on the Nodalities Blog at http://bit.ly/4qu13W

ESTC2009

December 2 -3, 2009 | Vienna, Austria

Attend Europe's most prominent and authoritative conference on progressing semantic technology markets.

European Semantic Technology Conference

brought to you by STI · INTERNATIONAL

a initiative

Become a certified semantic technologist.

Next course: Semantic Technology Specialist
November 23-27, 2009 | Vienna, Austria

semsphere

Register online: www.semsphere.com