Nodalities

THE MAGAZINE OF THE SEMANTIC WEB

talis®

www.talis.com/nodalities

Issue 14

INSIDE

Read Previous Issues of Nodalities Magazine Online

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

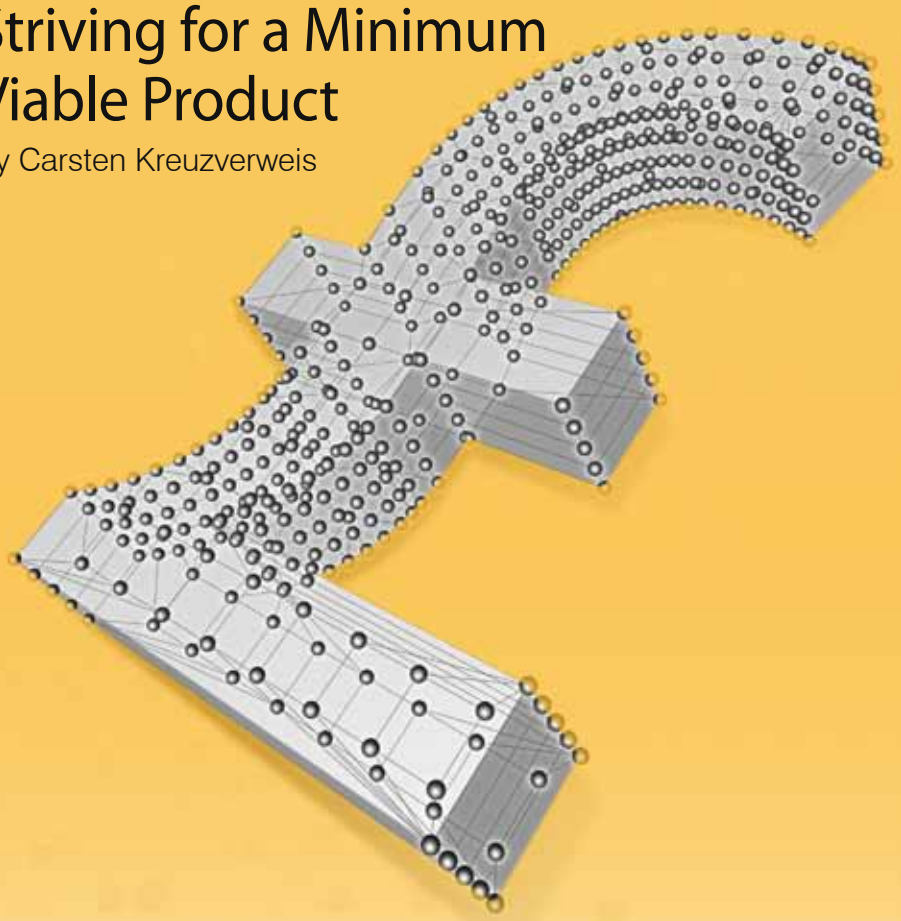
twitter

Follow Us On Twitter
@nodalities

Selling Semantics

Striving for a Minimum Viable Product

By Carsten Kreuzverweis



Semantic technologies are a great technology, but for the majority of people they are just that: technology. Selling semantic technologies to business people, investors, or end users is hard. Usually they don't seek semantics, they seek solutions to their problems. Recently we have started a spin-off company[1] of the University of Koblenz-Landau and successfully acquired a start-up funding from a German government programme[2]. While we control the technologies that we have been working with for several years now, the hardest endeavour so far has been to find the product. The problem was not, that we did not know who our customers are. We gained a lot of experiences in the media industry and knew that our technology could be employed there very well. The problem was actually to also convince the media industry of this fact.

continued on page 3

SUBSCRIBE NOW
Nodalities Magazine is made available, free of charge, in print and online. For back issues, please visit www.talis.com/nodalities

FOLLOW US
Follow Nodalities Magazine on Twitter: @nodalities

Editor's Notes

Nodalities has, for three years, told stories from around the ever-expanding sphere of Linked Data. As Editor, I have had the privilege of helping to tell those stories, to meet so many interesting people, and take part in the growth of the Linked Data community.

This will be Nodalities' final edition.

At Talis, we have never been a media or publishing company--we build data platforms and applications--but Nodalities was created to help our community grow and give a voice to those joining and building with us. The magazine has fulfilled it's purpose, and I am very happy for other media now to take responsibility for telling its stories.

As a final issue, I do not want to look back for context, but forward to new and emerging developments on the Web of Data. So, I have the added privilege of presenting a collection of stories with an emphasis on starting up for the final time.

If you would like to get in touch about any of these stories, or would like to catch up with me about what Talis is working on next, please feel free to contact me on zb@talis.com

Best,
Zach Beauvais

Nodalities Magazine is a Talis Publication

ISSN 1757-2592

Talis Group Limited
43 Temple Row ,Birmingham, B2 5LS
United Kingdom
Telephone (within the UK): 0870 400 5000
Telephone (outside the UK):+44 121 717 3500
www.talis.com/nodalities
nodalities-magazine@talis.com

Continued from front page.

By Carsten Kreuzverweis

In this article we outline the experiences we made on our way from our intial idea to a minimum viable product. As we have learned in the past 2 years, something that is theoretically great, is not of much use, when others don't understand why it is great. This article will therefore be hardly technical, but it is rather meant as a case study of how we start to sell semantics without mentioning the term "semantic" at all. Although we still are a young company and just about to enter the market, following the described approach has helped to do this now. If we followed our initial "vision", market entry would probably be still 1 year ahead.

The key idea comes from the lean startup community[3]: enter the market as soon as possible and learn as much as possible from customers. Most lean startups are less technology focused as we are, but rather rather sell more traditional products. They usually come with a business idea and then find the technology that fits them well. We actually did it the other way round. We had the technology and were required to find the business. In the following we will present our initial vision, discuss why this vision is hardly understandable to our potential customers, and show how we fixed this problem with our first product. We also give a short overview on the technical realisation, before we conclude this article.

Our Vision

Media management, e.g. managing images or videos, is an important task in many companies. TV stations, newspapers, and archives, among others, earn their money by producing, publishing, or selling media. Other companies require media for internal purposes, such as documentation or presentation.

A typical problem is to find a certain media asset in the enormous pool of media that is available. Most people search using keywords. Since the majority of media items do not contain text (or at least no text that can be indexed), annotations are crucial for any media management system. Most tools provide means to annotate media items using text or keywords. But the annotation is time-consuming, tedious, error-prone, and expensive.

Furthermore, most tools don't support the user in choosing the right keywords, which leads to increased effort required from the user and to a well known mismatch between the annotations that are provided and the terms that are searched for. A user annotating an image of the Brooklyn Bridge will probably choose a keyword based on what he sees in the image, such as "Brooklyn Bridge". However, someone searching the media archive might be looking for suspension bridges in the USA, because he is writing an article about this topic. The image might be a candidate for him, but it will not be found, because the annotator has used a keyword that the author of the article does not use for his search.

Our goal therefore is obvious: we simplify media management! We have the technology, we have gathered the expertise within several EU-funded projects, and we know what customers want. Everything started with Semaplorer[4], our contribution to and the winner of the 2008 Billion Triples Challenge. By mapping different datasets to images crawled from flickr, we showed a prototype that allowed for faceted browsing on flickr images, displayed images on a map, and provided background information. Semaplorer nicely presented the benefits of using linked data for media browsing.

Our initial vision was to extend this to a linked data based media management system. The idea was to radically change the way media is managed. Instead of using pure text, media items are linked to the concepts they contain or refer to. The benefits for our customers are manifold:

- Concepts carry a clear semantic meaning, they are not ambiguous.
- Links to other concepts and datasets "pull in" additional information. If you annotate an image with "Brooklyn Bridge", you implicitly annotate it with New York, USA, and suspension bridge. One action, four annotations.
- During search we can exploit all the information contained in the data, e.g., to compute powerful rankings, to provide faceted search and browsing, and to provide full-text search including synonyms. The concept USA for instance also provides labels with alternative names, such as "The States", "United States of America", and so on.
- Furthermore, once semantic annotation has been accepted, we can use reasoning techniques and rich domain models to further improve the annotation and provide for explicit semantics, provenance tracking, and inference of implicit information.

Given this list of benefits and given the knowledge of what semantic technologies can accomplish, it is obvious that they are superior to traditional technologies, and as a consequence, they are exactly what customers want to have.

Suddenly you don't use text, you use concepts. You don't change properties of media assets, but you link it to other information. You don't use text search, but now you browse with facets. Our customers are media professionals and not computer scientists, so they will probably not be able to see the benefits.

Reality

Presenting our vision to other people, including business angels, potential customers, and end users, however, confronted us with reality. Our vision did not make much sense to them, because they didn't understand and did not realize what our technology would change and how it would be improve media management. Basically this is due to following problems.

The first problem is that customers do not understand. And this is not even surprising. Someone using a Wordpress blog probably does not understand the relational calculus or the HTTP protocol, and fortunately he doesn't have to. The same problem holds for Linked Media. The customer does not understand what semantic technologies can do. We have to convey to him the benefits using terms and examples he is familiar with.

The second problem is that the vision is too complex. Our vision was supposed to revolutionize media management and to be a radically new way of managing media assets. And that is just too much. Suddenly you don't use text, you use concepts. You don't change properties of media assets, but you link it to other information. You don't use text search, but now you browse with facets. Our customers are media professionals and not computer scientists, so they will probably not be able to see the benefits.

And finally, this vision just comes too early. A customer, who does not even have good annotations yet, can not use Linked Media. His problem right now is, that his text annotations are too expensive and do not

work as expected. His problem right now is that an image annotated with "Brooklyn Bridge" is not found by a search for images depicting suspension bridges in the USA. The customer wants a solution to that problem, but not a radical new way of managing media. To him, radically new means probably expensive, and expensive in that case is rather contra-productive for him.

In our case, we identified the major pain point of our customers: the annotation itself. It is expensive, time-consuming and tedious, and it is error-prone due to the mismatch between what the annotator sees and what someone searching might be looking for. However, it is crucial to our customers business.

A Minimum Viable Product

The core question to answer therefore is: how can we simplify our idea in order to achieve something that everyone, and especially the customer, understands. Our approach was to define a Minimum Viable Product (MVP). An MVP is a great means to reduce your vision to a product that can be realized, presented to a customer, and sold. Since it is minimal, it is typically cheap to realize. So even when the assumptions made were wrong, you have not wasted too much resources and can easily create a modified product and try again. It enables you to test your ideas directly in the market with very small effort.

In our case, we identified the major pain point of our customers: the annotation itself. It is expensive, time-consuming and tedious, and it is error-prone due to the mismatch between what the annotator sees and what someone searching might be looking for. However, it is crucial to our customers business.

The core problem with annotation is not even the fact that keywords are used instead of concepts, but that either not sufficiently many or not the right keywords are used. Given the right keywords and appropriate quality, a traditional keyword based system would work much better. Therefore, instead of revolutionizing media management, we start to simplify the annotation.

What can be done to improve annotation? We reduced this to two things: supporting the user in typing the right keyword and proposing additional keywords. The functionality is implemented as a web service that is capable of providing auto-completion for keywords and that provides keyword proposals based on a list of existing keywords[5]. Secondly, we implemented a small add-on to Apple Aperture, a (semi-) professional image management software for the Mac. This app is called Annotator for Aperture®[6], and is soon to be released in the Mac App Store. The web service enables us to sell our product as Software-as-a-Service in a B2B market and the Mac App is a valuable tool for us to a) demonstrate what the service does, and b) test it in the market. In the following, we will give an overview of the technical details, and show that although the product looks (intentionally) simple, it is backed by a rather sophisticated infrastructure.

Technology

Our core product is the web service, which we offer as Software-as-a-Service. For such a service we need high scalability. For this reason, everything is deployed within a cloud infrastructure based on Eucalyptus and thus Amazon AWS APIs. We consequently apply RESTful principles for all modules internally. Basically, none of the modules of our system share any state, but only exchange representations. Any module can therefore be deployed on another instance and its functionality can be exposed using RESTful HTTP APIs. We use datasets available as RDF, and store data in a Sesame 3 backend, that was extended with custom functionality. Fulltext search is realized using the Lucene Sail. Distributed access is currently reimplemented based on the Semaplorer backend. We have implemented an infrastructure for transformation and mapping of datasets to a common representation and use different technologies for ranking resources both globally and in the near future also locally, i.e., on a per request basis. Everything is implemented in Scala and deployed within an OSGi runtime environment (Eclipse Equinox). Two core aspects for the infrastructure are to be flexible and scalable. The former provides as the means to easily deploy new products or enter new markets, the latter enables us to provide SaaS. Both allow us to integrate with many different products, which is a core requirement in a market which is very heterogeneous with respect to the deployed systems.

While it was possible to define a very minimal product, the infrastructure to provide this product is rather sophisticated. This is one of the differences to other lean startups, where often the technology plays a secondary role. For us, however, the technology is a key aspect, and this technology is pretty complex. The important part is, that an application like Annotator for Aperture® completely hides this complexity from the user.

Conclusions

We are still at the very beginning and just started distributing our services and our app. While we have not yet sold our MVP a million times, we have achieved two things: a) we have a product and b) we are able to explain customers what we provide. If we followed our initial plan, namely revolutionizing media management, we would have neither. For us, the trick to sell semantics was to not sell semantics. Now we are selling the solution. We started from what customers use today and what they understand.

We hope this article gave you an idea of how we started to sell semantic technologies without mentioning semantics at all. Semantics may be crucial to provide the actual services and to provide a long term vision, but the product should be defined by features that customers understand and value.

References:

[1] http://kreuzverweis.com/

[2] http://www.exist.de/englische_v ersion/index.php

[3] http://leanstartup.pbworks.com/w/page/15765221/FrontPage

[4] Schenk, S.; Saathoff, C.; Staab, S. & Scherp, A. (2009), 'SemaPlorer -Interactive Semantic Exploration of Data and Media based on a Federated Cloud Infrastructure', Journal of Web Semantics.

[5] http://kreuzverweis.com/annotation-webservice/

[6] http://kreuzverweis.com/annotator/

Actionable Linked Data with SPIN

By Holger Knublauch, VP of Research and Development at TopQuadrant. He introduces SPIN, a new W3C Member Submission that utilizes SPARQL to enhance linked data models with rules, constraints and functions, similar to how object-oriented languages link classes to actionable behaviour.

The current family of modelling languages for the Semantic Web defines shared vocabularies that are great for describing static aspects of data in terms of classes, properties and instances. For example, SKOS can be used to define concepts and their relationships with each other. OWL makes it possible to define classes and their associated properties. And then there are thousands of emerging or established ontologies covering domains such as people, companies, products, etc. All this is great to categorize data that is being published online, or data that lives in internal data silos or spreadsheets.

One of the strengths of the Semantic Web language stack is that data from many different sources can be easily linked together to gain integrated views. Another strength is that Linked Data is inherently self-describing, which means that a software agent can look up the definition of an object's type by following a URI. Putting those advantages together means that it becomes possible to utilize generic data browsers, generic mash-up tools and generic search utilities that know almost nothing about specific domain models, but "discover" all this information on the fly.

However, with the current stack of ontology languages, the ability to make sense of Linked Data is somewhat limited. The SPARQL Inferencing Notation (SPIN) extends the expressivity of RDF-based languages by adding a layer of constraints, rules and functions that make Linked Data actionable and more self-describing. SPIN has been developed at TopQuadrant over the last few years and been tested and improved through constant user feedback as part of the TopBraid product line. Since SPIN has become sufficiently stable, we believed the time was right to share it with the rest of the Semantic Web community. SPIN was published as a W3C Member Submission in April 2011, which means that other tool vendors are encouraged to include SPIN-based features into their products. Furthermore, SPIN may drive the evolution of related W3C standards in the future.

So what can SPIN do for you? Let me start with a simple example. Let's say we publish a model about geometric shapes, including rectangles and squares. It is straight forward to define classes such as Rectangle, with properties such as width and height in OWL or RDFS. SPIN defines a couple of very simple properties that can be used to link those ontology classes with SPARQL queries:

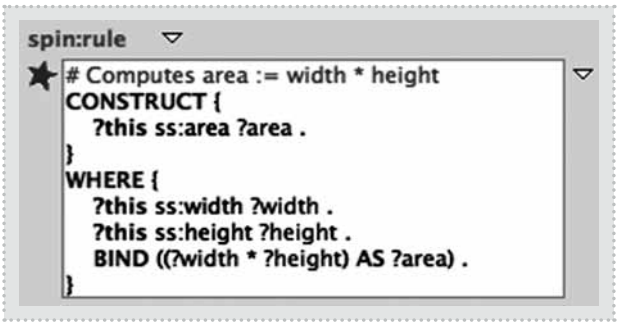
- The property spin:constraint can be used to link a class with a SPARQL ASK or CONSTRUCT query that can be used to verify integrity constraints. For example, you can state that in a square, width and height must be equal.
- The property spin:rule links a class with a SPARQL CONSTRUCT query or INSERT/DELETE request that defines how new values can be derived from existing values. For example, you can state that a Rectangle's area is the mathematical product of width and height.
- Finally, the property spin:constructor can be used to initialize new instances with suitable (default) values.
- The following TopBraid Composer screenshot illustrates how those SPIN constructs can be edited and maintained. The image shows a SPIN rule on the form for the class Rectangle. SPIN uses an object-oriented approach, in which instances of the associated class are accessed via the variable "?this".

From a technical point of view, the linkage between classes and those constraints and rules uses an RDF-based representation of SPARQL. This means that the queries are not just stored as a string, but as a rich data structure similar to how OWL restrictions and SWRL rules are stored. As a result it becomes easier to embed and correlate SPIN queries with the domain models that they belong to. For execution and display purposes, those RDF structures are converted back to SPARQL strings, and this way they can be executed by any RDF store with SPARQL support.

If a SPIN-aware Linked Data engine finds those constraint and rule definitions, it is able to drive user interfaces or data processing tasks with them. For example, data entry and search forms can automatically reject values that violate constraints. Rules can be fired to automatically calculate some property values from others. Rules are often used to define transformations, and these transformations benefit from the rich expressivity of SPARQL with all its built-in functions and constructs to filter, bind, aggregate and join data. So instead of just seeing a bunch of names for classes, properties and instances, Linked Data engines can use SPIN to do more with the data that is being explored.

One of the strengths of the Semantic Web language stack is that data from many different sources can be easily linked together to gain integrated views.

In addition to constraints and rules, SPIN also introduces a new concept into the Semantic Web space: user-defined functions and templates. Like with any object model, users often want to ask the same queries again and again. In the case of programming languages like Java, functions such as getArea() can be designed to conveniently compute the area of a rectangle. Instead of having to re-write the formula time and again, programmers simply call the function by its name and don't have to worry about the internal details of its implementation. SPIN uses a very similar idea,



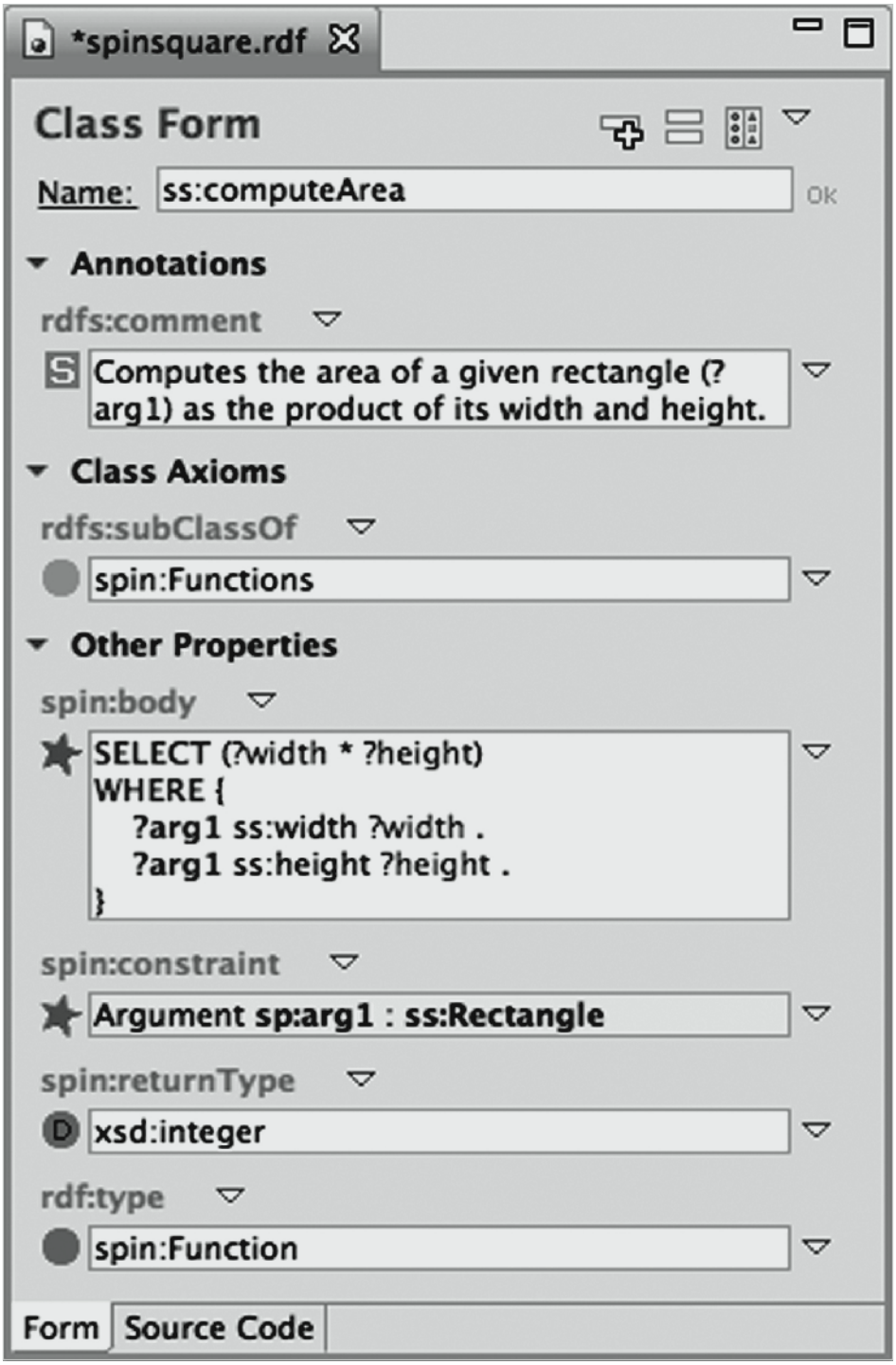
and provides an RDF-based vocabulary to define new SPARQL functions. Each function defines a formal description of its arguments, its return value, and a body query that shall be executed whenever the function is called. For example, a function to compute the area of a Rectangle may take the Rectangle as its argument, then runs a SPARQL SELECT query to get the rectangle's width and height and finally returns the product of those two values. Such a complete function definition is shown in the following TopBraid Composer screenshot. (see image right)

Such SPIN functions can be shared on the Semantic Web in a purely declarative form. If a SPIN-aware SPARQL processor finds a query that uses this function, it can look up the function's definition by following its URI. This way, libraries of reusable building blocks can be shared, leading to a new paradigm in web-oriented software development.

A final aspect of SPIN that is worth mentioning here (without going into details) is the concept of templates. A SPIN template is a reusable SPARQL query that has been given a URI similar to how SPIN functions are declared. The queries in such templates can leave certain values "blank" using a SPARQL variable, and users of the templates can fill in those blanks for specific use cases. This makes it possible to reuse the same query in multiple places and allows for the creation of higher level languages for users who are not familiar with SPARQL. A simple example for a SPIN template is a constraint check that all values of a given property must be greater than zero. The template leaves the particular property blank, and ontology designers can reuse this template filled in with their specific properties. A fine example of such a SPIN template library has recently been published for Semantic Web Data Quality.

While I could only scratch the surface of SPIN here, I hope I was able to illustrate some of the power that SPIN brings into the Linked Data universe. In addition to the static structure of data, SPIN can add behavioural aspects, and these strengthen the very foundations of the Semantic Web: self-describing linkable data.

Judging from our user community, SPIN is rapidly gaining momentum. It is relatively easy to learn and can be efficiently executed on any RDF data base with SPARQL support. If you want to learn more about SPIN, the following links may be helpful:

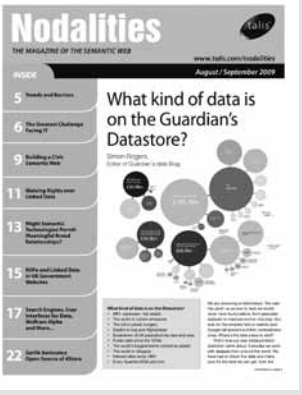


SPIN W3C Member Submission http://www.w3.org/Submission/2011/SUBM-spin-overview-20110222/

SPIN Community Site with many links and articles http://spinrdf.org/

SPIN Support in TopBraid (Free Edition) http://www.topquadrant.com/products/SPIN.html

SPIN API (Open Source, based on Jena) http://topbraid.org/spin/api/



Nodalities Magazine

Nodalities Magazine is made available, free of charge, in print and online.Back issues are available at www.talis.com/nodalities or email nodalities-magazine@talis.com.

Open Data in Italy: why we look to the Save the Data Campaign

By Matteo Brunati, Social Media Strategist at Metodo S.p.A and TeamSystem Group, Italy. Contributor to the Open Data movement in Italy and the RDF W3C Working Group.

Two years ago the US Administration launched the data.gov site, thanks to the famous Obama memorandum on transparency and government data accessibility, [0] which mainly encourages mainly three values which are associated with a new idea of government: transparency, participation and collaboration. It was one of the first interesting actions of the Obama administration, and one of the first basic actions inside an official government context, which speaks using the typical language of the Net culture of sharing and participation.

Thanks to this memorandum, the US federal budget enables, along these years, not only data.gov, but also usaspending.gov with the IT dashboard (it.usaspending.gov) and other sites which work as enablers for the new "Government as Platform" mantra.

For countries like Italy with cultural gaps on topics, especially inside an official government context, on topics concerning the digital media, it was like a boat covered in darkness lightened by a lighthouse.

As a follower of the Semantic Web community, it's a pleasure to witness how these technologies can help to achieve goals proposed by Obama's memorandum, because this context can help to a better spread of the Semantic Web and Linked Data technologies in the global community of developers and IT decision makers.

Now, in the middle of a better global understanding of the connections between the vision of Open Government and Open Data movements across the globe, there is a stop to the government support for these examples of a better democracy process: probably cut off of the federal budget from $37 million to $2 million. It seems that the last offer is increasing the budget to $8 million: but it's far from the actual $37 expense, so the issue remains opened.[1]

What the "Save the Data campaign" is about

Obviously US citizens move on to make the Congress more aware of the importance of not cutting off the budget to support these transparency data sites. One of the main actors in that context is the Sunlight Foundation, that started an interesting bottom up action, called "Save the Data" campaign.[2] It helps the citizen to sign in a letter to the Congress leaders asking them to protect funding for the Electronic Government Fund, and to contact them directly and finally to spread the word using social media.

At the time I'm writing this article, more than six thousand people signed the letter to the Congress, hundreds more have written letters, called their representatives, written blog posts or shared the campaign with their friends.

But this not only happens in the USA, also in Europe and Italy people talk about this campaign and help to spread the word.[3]

> *In a country where television is the main focus media to get informed about what happened around us, it's a long journey to make the people more involved with the culture and tools enabled by the Net.*

To make sure it's more understandable we have to avoid that people who are not involved in this studies should understand that the closing of data.gov imply the closing of the Open Government at all. It's not that straightforward, luckily.

A story of the Open Data movement in Italy

I'm not a US citizen, but as a passionate user of data, and as a believer on the Web as a main actor of future enabled by transparency, and data accountability, I'm worried about what is happening in the US.

As a matter of fact, in Italy we are fighting for a more credible way to enable a similar type of transparency regarding the government activities, using US data.gov and similar sites as examples of this potential. This potential is expressed using the availability and the accessibility of the raw data, collected in one central point, and this made it easier to seek for everyone.

In a country where television is the main focus media to get informed about what happened around us, it's a long journey to make the people more involved with the culture and tools enabled by the Net.

We have some difficulties to make the usage of Net and the usage of Web more understandable not only as a media channel, but also as a place and a platform of development, which can help us to enable a new augmented social dimension.

We know that we are not an example of Open Government as a platform (maybe as services available online), neither an example of a country leader in making IT technologies or leader in making an openness culture. Things like Creative Commons, Open Source and Open Linked Data are not known by the majority of people, also in our IT community.

In this context of cultural and digital division, not all things are bad; perhaps we can see the light at the end of the darkness. Italy is a living ecosystem as every country in the world. We made Italy, we can change it, and we can help to make it better. Young people are very smart, working hard to take the leap possible.

One year ago a small conversation was born in some blogs regarding Open Data movement and Semantic Web technologies. Also some articles were published in a special number of Nòva24, the technological insert of one of the main newspapers in Italy, called 'Il Sole24Ore', with some references published also in its online version, called 'Nova Review'[4].

During the summer of 2010, a tiny but strong bottom up discussion came up, mainly in Italian blogsphere: we called it "Spaghetti Open Data". Basically, we started with a lot of mails and some blog posts around the Open Data stuff, then we used a mailing list to a better support of our collective conversation. The clear thing is that in this conversation there aren't only geeks, hackers, or Open Data and Linked Data enthusiasts, as happened in a similar movements around the globe but there are also some people working directly in the Public Sector, in Rome, who want to enable

more transparency and citizen participation using a bottom up approach. They don't know anything about technical factors behind the Open Data movement, but they are curious and interested in it. In those months, the only official participant from Public Sector involved in Open Data activities was Regione Piemonte, an European leader and innovator in local open data policy development and practices. In May 2010, Piemonte launched, as a matter of fact, an online beta data portal - dati.piemonte.it.[5]

The second step of Open Data movement was a meme started from Alberto Cottica's blog, at the end of September 2010. This meme ended up in November of the same year with a launch of a small example of what a data.gov.it site could be in the future.[6]

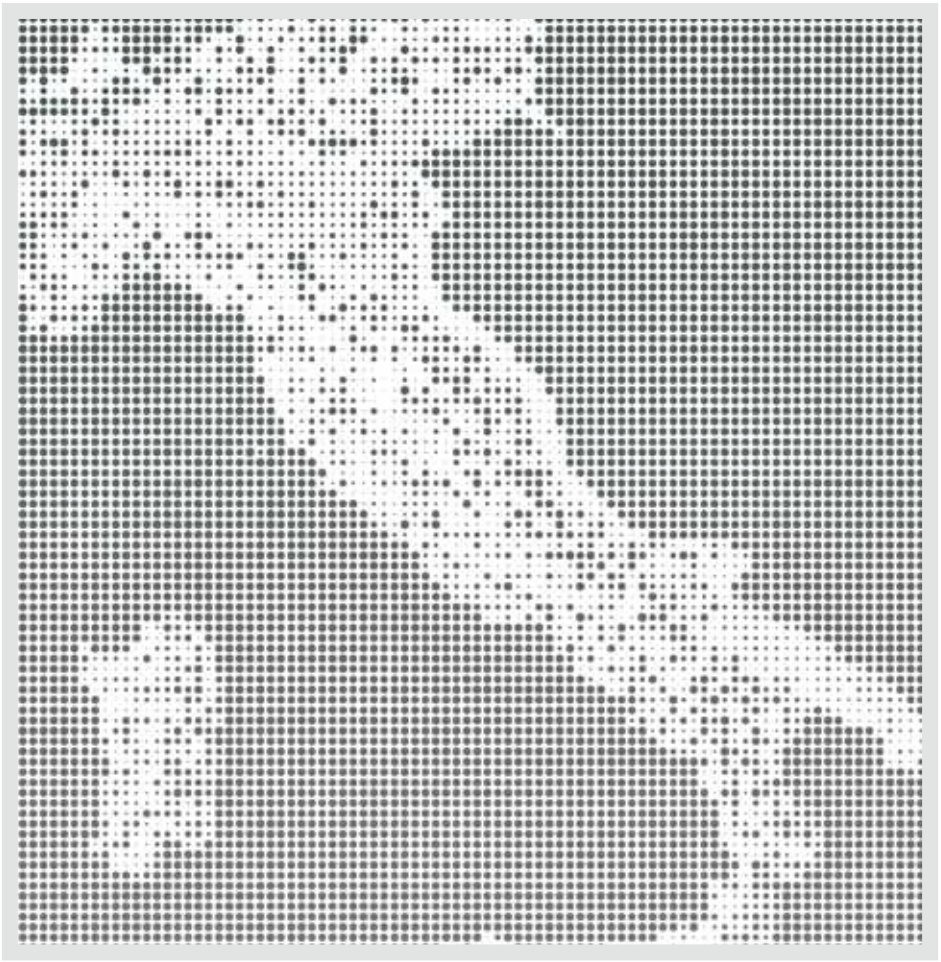
We thought that it would be nice to collect links related to public data available online in Italian government websites in one place. This was a kind of one stop shop for people interested in transparency not just in theory, but in the practice of extracting information from public data.

So, in November 2010 the Spaghetti Open Data site was born with 32 aggregated databases (only linked, not hosted). Not bad when you consider that data.gov, with all the firepower of the Obama administration, had 47 at launch. These datasets available aren't really Open Data, they are only public data, except a very few of them. The difference between public data and Open Data at that time wasn't so clear to people not aware of Semantic Web context. So at the beginning we collected both of them and then the community and the conversation made that point clearer, naturally.

In the same days the first edition of the 5-stars Linked Data scheme[7] was published. We used this scheme to make the overall context more understandable, using a column named "Reusability"[7.5], on the page showing all the datasets available.

The possibility of reusing data is one of the key concepts concerning Open Data. This data must be available for humans and for machines and must be shared with a legal license that allows to use them again, an aspect which has not been very well understood by data's publishers.

We don't want to reinvent the wheel: without money we made spaghettiopendata.org using the famous Simile Exhibit framework connected with some Google SpreadSheets, and nothing else. If some sites would have published data using RDFa standard or similar, we could have inserted them in a moment without any added cost.



Meanwhile, we were ready to make it better for a complete integration with Italian CKAN repository, it.ckan.net, same data back-end that powered also data.gov.uk. Because of the lack of time in 2010, we are working for it nowadays. We could maintain spaghettiopendata site as another visualization of the main it.ckan.net repository, thanks to API and SPARQL endpoints.

And, last but not least, we can improve it as quickly as we can.

In last months of 2010, movement has grew up, with the birth of two different associations related to Open Data. The first was Datagov.it[9] and the other one another was linkedopendata.it [10], born on January 2011. Datagov.it is a social platform open to collaboration in defining the OpenData Manifesto in Italy. Linkedopendata.it is a more structured evolution of our idea of aggregating Open Data in a Linked way, with Semantic Web technologies and SPARQL endpoints available for the users.

Things were keeping on move: two different meet-ups in Rome from September to December 2010, where other interesting ideas and apps came up, to enable the level of applications over data of Linked Data cloud.[11]

One interesting thing was the official role of the main Italian statistical agency, ISTAT: an important supporter of the movement, also with practical contributes as a Wordpress plug-in for its Data Warehouse ecosystem.

Conferences like FammiSapere.info[12] and

"La Politica della Trasparenza e dei Dati Aperti"[13] -with international people like Jonathan Gray of the Open Knowledge Foundation- are, nowadays, helping us reaching more mass media coverage and spreading the word.

From a regional level to an institutional one, our goal is to put Open Data movement on the agenda setting of our political class and mass media channels.

The challenge for Italy is to enable this level of institutional support at the regional level, because here we can make practical steps to move on to Rome with a solid proof of concept.

So, with this context in Italy, turning back in the US, what is our perspective about the "Save the Data" campaign? Why is it so important for us? Let's see.

It's not about data, is about transactional costs connected with data

The choice of closing data.gov is not only a problem of data which totally disappeared from Web. As an Italian start-upper said[14], "Start-ups will save Open Data".

I think it's a correct point of view: there are a lot of data start-ups in US which can store datasets, so it's not a problem of data availability. It's a problem of process, I think.

> *"Data.gov also plays crucial behind the scenes role, setting standards for open data and helping individual departments and agencies live up to those standards. Data.gov establishes a standard, cross-agency process for publishing raw datasets. The program gives agencies clear guidance on the mechanics and requirements to release each new dataset online. "*

This quote is taken from Harlan Yu's blog[15], and I totally agree with it. Italy is a country where all the stuff connected with Open Data value chain has a cost, not only related to money, but also among social and political efforts. So, every added step is an obstacle for the birth of data infrastructure. And not only in Italy, but in all the places where Open Data movements are young.

Without a political support, we have to improve our work with data: seeking for it around gov sites and making it easier to access.

Now, in April 2011, asking for support of an Italian politician outside the data.gov story is really difficult. Making an official Italian data.gov while US is closing its own one, makes a politician sceptical about all Open Data value chain.

How to reply to this observation? If a start-up could easily find the data to work with, it could focus its own investments on the mash-up idea or the service. But if the data has to be aggregated and made usable, the total cost is higher and a start-up could be afraid of this high cost.

Another cost for a potential Open Data start-up is the lack of standard published information, because the data must be transformed again.

So without an easily access to data, a standard to publish them and a strong sharing process for gov data publishers, the entire Open Data movement in a country like Italy is at risk of imploding on itself: that's why we are working on bottom up guidelines for Public Sector's actors, like 'Come Si Fa Open Data – Versione 1.0'[16].

Save the Data in the US means save the importance of sharing process

As we can see, it is a long journey to enable political support on these things, and a US leap backwards can damage all work done so far. It's not a technical problem: we have to focus on the process and its institutional approval across Public Sector agencies. The enabler factor is here, and cutting off symbols of Open Data as data.gov breaks out our official references, and so our credibility for our local government.

Notes

[0] - http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment/

[1] - http://www.guardian.co.uk/news/datablog/2011/apr/05/data-gov-crisis-obama

[2] - http://sunlightfoundation.com/savethedata/

[3] - http://www.apogeonline.com/webzine/2011/04/05/save-the-data-sfumano-i-finanziamenti-di-obama

[4] - http://novareview.ilsole24ore.com/pubblica-amministrazione/

[5] - http://www.epsiplus.net/news/news/italian_piemonte_region_european_opendata_leader_and_innovator

[6] - http://opensource.com/government/10/11/spaghetti-open-data-little-thing-feels-right

[7] - http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

[8] - http://www.spaghettiopendata.org/dati

[9] - http://www.datagov.it

[10] - http://www.linkedopendata.it

[11] - http://www.meetup.com/The-Rome-Semantic-Web-Meetup-Group/events/past/

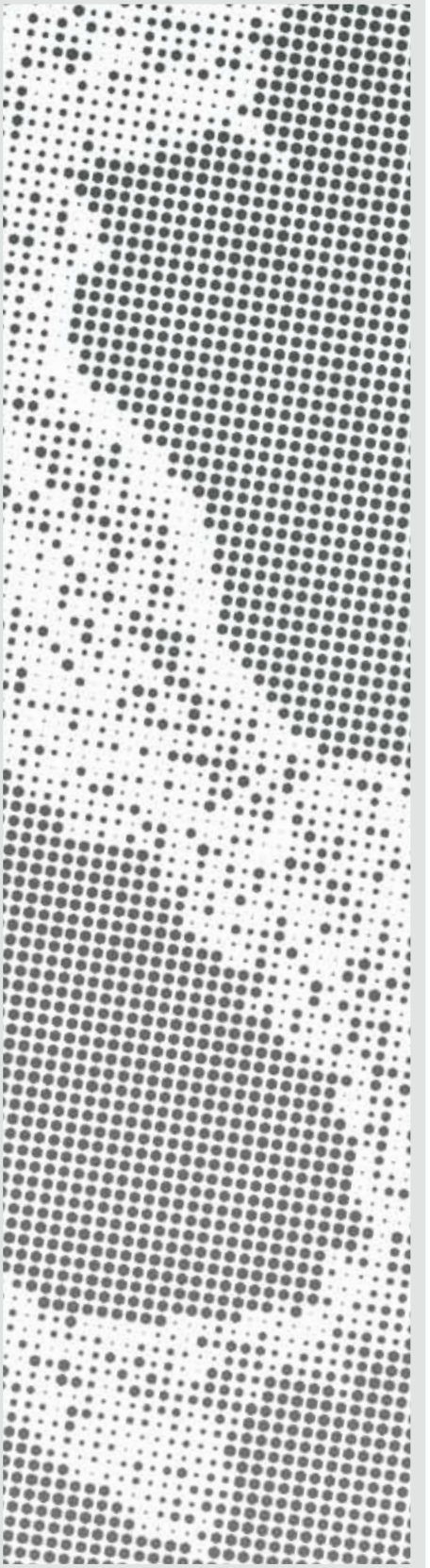
[12] - http://www.fammisapere.info

[13] - http://www.linkedopendata.it/la-politica-della-trasparenza-e-dei-dati-aperti

[14] - http://www.linneapassaler.it/2011/04/07/startups-will-save-open-data/

[15] - http://www.freedom-to-tinker.com/blog/harlanyu/what-we-lose-if-we-lose-datagov

[16] - http://www.scribd.com/doc/52509290/Come-Si-Fa-Open-Data-Versione-1-0



OpenStreetMap

By Tom Chance, OpenStreetMapper

The first chink of open data light for the OpenStreetMap community in the UK was a negotiated release of the NAPTAN database, flooding the UK map with bus stops to much delight.

But no sooner had we carefully imported it, one local authority at a time, than mappers started to notice errors. Some bus stops weren't in the right place; they had the wrong name; or they no longer existed. We also wanted more information about them, such as whether they have a bus shelter and which routes stop at them.

So began a process of checking and enhancing the open data.

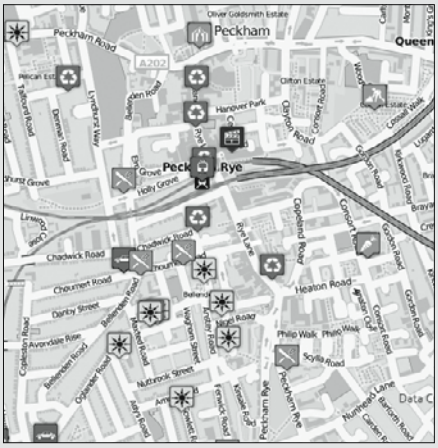
With a few small exceptions, none of this has been fed back to the Department for Transport. Some mappers have shared their results with their local authority transport team, but that's about it.

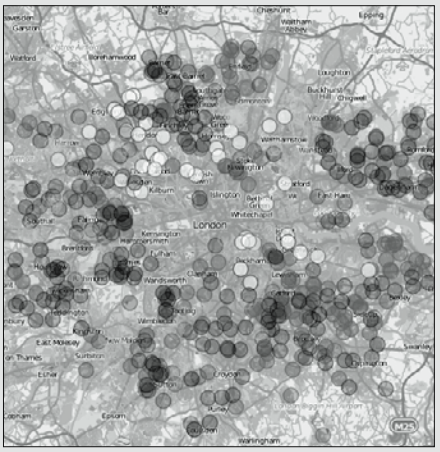
I had a similarly mixed open data experience looking at open data the Greater London Authority released on allotments. The data was collected from boroughs by an Assembly committee a few years back, and while more comprehensive than our crowdsourced data it only had midpoints rather than the shape of each allotment, and included allotments that no longer existed. It also didn't have any community growing spaces.

These frustrations were the backdrop when I started working on OpenEcoMaps, an eco/green/sustainability spin-off of OpenStreetMap. OpenStreetMap is a community project trying to map the whole world and share the data under a free/open license. Unlike Google Mapmaker, the focus is open data rather than freely donating your sweat to a company.

My first motivation for OpenEcoMaps was simply to make all that wonderful open data we had hidden behind technical barriers available to local green groups who love to put maps on their web sites and street stalls. Geeks all too often forget just how skilled they are.

It's no surprise that OpenStreetMap volunteers focus on entering data and providing tools that other geeks can use to build fantastic apps. The conventional wisdom in the open data world seems to be that geeks are the gateway for the remaining 99% of the population by creating consumer apps and visualisations. I wanted to give that 99% a way to contribute to and use our data on eco features without needing to learn about APIs and shapefiles.





OpenEcoMaps takes data about "eco" features stored in OpenStreetMap and turns them into KML files that can be shown as overlays on a map, making it easy for people to find out where they can get a vegetarian meal, forage some wild fruit, spot a solar panel, recycle a can, pick up a car club car, or spend some money in a cinema.

People can use the map on the OpenEcoMaps web site; embed that map on their own web site; stick the KML files into a map they already have; or even explore them in Google Earth and other GIS software.

OpenEcoMaps also offers a customised version of the new editor, Potlatch 2, which has all the features shown on the maps as presets. So, having created an OpenStreetMap account, people can simply drag and drop a new wind turbine onto the map or draw in a new recycling centre.

While I am no programmer by training – I work as a policy researcher at the GLA for Green politicians – I have gradually been hacking the code into shape and it is trivial to add new packs of KML files for different towns. At the time of writing we have maps for London, Glasgow, Reading and Exeter.

My second motivation was to explore how we could start collaborating with organisations who were starting to share dumps of open data, using it to complement crowd-sourced data. I have been looking at open food and energy data as starting points.

I got onto energy data at the request of a friend who runs a sustainable energy social enterprise. They wanted to show people just how many renewables there already are in our local Peckham to inspire people and show it's already main-stream, so we went out to map them. I also asked the council, the Greater London Authority and the Department for Energy and Climate Change what they knew. We discovered that nobody really has a clue!

For the whole of London we got about 3 renewable generators from DECC, 41 fire stations with renewables from the GLA and – to date – nothing of use from Southwark Council. We added in a dozen or so solar panels and wood boilers we could spot in our local streets and owned by people we know, and continue to look at ways of working with companies and the public sector to improve our knowledge of what's actually out there.

The food work came from looking at allotments. There's a lot of interest in food in London at the moment, but again nobody really knows how much space is being used to grow it. So I'm talking with the GLA, Sustain, Groundwork, Southwark Council and local enthusiasts about using OpenStreetMap as the platform for collaboratively mapping and sharing that data.

If you're interested in OpenEcoMaps, I could really do with some proper programmers and I'm always interested in discussing new ideas for data collaboration.

www.openecomaps.co.uk

tom@acrewoods.net

The RDF platform that grows along with you

By Ben Lavender, Co-founder of Datagraph

Dydra is the product of a few years of investigation and research by myself and my co-founders. We hope that Dydra plays a role in the growing adoption of RDF for commercial purposes by providing a turnkey RDF store-as-a-service. What follows is an explanation of why we feel it's essential to the future of the semantic web to provide that service together with the Ruby RDF infrastructure which we developed over the course of the last year.

Knowing where we're coming from will help clarify the problem we're trying to solve. On the one hand, we're very much semweb nerds, but we're also part of the startup community. We watch the 500 Startups demo day livestream and make accounts with all the Y Combinator companies to see what's happening. And nobody there was using RDF or taking advantage of the growing corpus of linked data. Not only that, these entrepreneurial communities are big adopters of non-relational database solutions like MongoDB, Cassandra, and others. These are people who are widely dissatisfied with existing relational databases. And yet none of them are picking a non-relational solution with a standard behind it, such as RDF in particular.

I find this quite frustrating. In 2008, I had my first real taste of RDF and the semantic web. The project was a far-reaching project using Drupal, with the kind of complicated cross-site metadata problems that tend to lead people to reinvent RDF, but we were lucky enough to have someone on the team who had used it before. After figuring out what it really does, I jumped into it with both feet and never looked back. I'm still using it every day. That RDF has so little traction in the startup community irks me.

After our work in Drupal, a we found a number of opportunities in publishing data with RDF, but we were more interested in consuming and using the vast amounts of data already out there (and besides, the project to output RDFa by default in Drupal was already in capable hands; we take no credit for that work). We dove in to other projects, ones with interesting problems that we thought RDF might help solve. Those projects lasted about a year and a half, and we were still looking for them as of about 9 months ago.

And yet, the project was never quite right to use RDF. We kept heading back to a more agile toolset, despite knowing that if these projects were successful, we'd be wishing we'd used RDF sooner rather than later. It seems that every project of sufficient complexity eventually re-invents half of it, badly, and we wanted to avoid falling in to that trap. Again.

To be fair, RDF is never going to displace the relational database. And the first database many projects need is just that. But as someone who understands the power of RDF, I'm mystified whenever I see startups blogging about recommendation algorithms and social network traversals in application code above the relational database. The startup world is using Ruby on Rails plugins that actually store cached copies of common graph traversal algorithms in key-value stores! These problems are pleading to be modeled better. So why aren't they?

In 2010, we felt that part of the problem, at least in the startup world, was a lack of agile tooling. This is a world that has largely left Java behind in favor of Ruby, PHP, and Python, and the Ruby RDF community was not yet unified under a common banner. So we wrote an entire suite of RDF libraries in Ruby, called RDF.rb. This is tooling we still use and are proud of. This solved a lot of problems, and RDF users and Rubyists both are coming to Ruby, which we're quite glad to see.

IKnowing where we're coming from will help clarify the problem we're trying to solve. On the one hand, we're very much semweb nerds, but we're also part of the startup community.

But early last year, we had an opportunity that opened our eyes to some other problems. We were able to evaluate most of the big-name RDF stores on extremely beefy hardware; 24-core, hundreds-of-gigs-of-RAM hardware. We spent plenty of time just trying to get them to run, and performance was disappointing; it was hard to make these products make effective use of all the resources available to us. We were forced to realize that library tooling was not going to be enough; we'd be hard pressed to recommend using RDF to our clients, or to make the decision to start a business based on it, with what we were using. Things were just too complicated to get real work done.

So we thought about what we'd want in an RDF store that we were missing, and it came down to an ease of use. The performance and features are already there for most use cases. This is a known problem, and a number of solutions exist, with well-known projects at all levels, from full-fledged installed databases to embeddable C libraries. But these aren't what lean startups want anymore. Services are the future, cloud services in particular. The gains in efficiency are too great to ignore.

For example, Amazon Web Services provides a number of cloud database solutions. These are not just traditional relational databases, but some kinds of other databases as well. Using them consists of turning them on, then pointing an application at them. It takes seconds. For RDF to succeed with startups, it needs something that fast and that easy, and it needs to exist in the same cloud that runs most startups' infrastructure anyway, which is to say Amazon's.

Dydra answers that need. Dydra is an RDF store delivered as a service, with lots of attention paid to the interface that software developers see. It's dirt-simple to use, and it allocates resources transparently to the user--no concerns about how big the backing systems are. If you're trying to start a business with less than $15,000 in funding, Dydra's aim is to let you solve your problems with RDF and be under budget.

We believe that filling this hole in the market can help move RDF forward in a big way, but that's not all. You may have noticed that this article talks very much about RDF, and not much about Linked Data (though that's about to change). We believe that this unmet need is a big reason that Linked Data still faces challenges, despite it's phenomenal growth over the last couple of years. (A quick disclaimer: the rest of this piece is going to talk as if all Linked Data is RDF, which is manifestly not the case. We feel, however, that RDF's universal identifiers are a hugely important piece of the Linked Data puzzle, and that real success in the area will come via RDF. We're not out to minimize the

impact or effectiveness of the other high-quality datasets out there, but this is what we work with.)

We believe Linked Data is challenged by the lack of a Dydra-like market participant in two ways. First, RDF being difficult to use does not predispose a Linked Data consumer to use RDF directly. Second, operational and administrative concerns prevent wide-scale adoption of Linked Data directly.

How does complexity in using RDF for one's own data do harm to the Linked Data ecosystem? Well, Linked Data, in whatever format it may appear, gains its benefits from an interacting ecosystem of data providers. The large-scale, interacting ecosystem of benefits envisioned (correctly) by Linked Data proponents won't be realized until RDF is worth a programmer's time to use for their own data. Until then, it's just another random, published format (or formats) to be downloaded, parsed, and turned in to something else to be used internally by an application. RDF is just another problem to be solved until the problems with using RDF are solved.

The end result? Consumers of Linked Data won't be publishing any. The bazaar-like ecosystem that has been so beneficial for open-source software can't get off the ground until it's trivial for consumers to push their changes upstream. Until those consumers are happy to use RDF for their own data, and thus bring in public data as if it were their own, Linked Data won't take off. Instead, the carefully modeled and published Linked Data will be locked away into innumerable one-off private formats when it's used, never to return to the community.

The second way we hope Dydra can help Linked Data is by providing a local cache of it. As mentioned before, Linked Data has administrative and operational concerns that go far beyond the technical capabilities of RDF. Namely, any real application cannot use data that is just *linked* to. An application providing a service to users will need to make a local copy so that it can control the factors important to the value it provides, such as performance characteristics and availability.

Consider a hypothetical example to demonstrate the administrative issues around using Linked Data. A budding entrepreneur thinks up a valuable, innovative application of an existing, published dataset. He takes his idea to an investor. "I'd like to use a publicly available SPARQL endpoint of this widget dataset to make an awesome website mashing up widgets and wobbles". The investor, diligently investigating his prospect, inquires as to the nature of this SPARQL endpoint. "It's run by some extremely competent folks at a technical university in Germany", he replies. At which point the conversation ends, and no investment occurs. No reasonable business can be built relying on a public data access point, regardless of the technical capabilities of the people running it. What if they decide to take it down, have a planned outage, or remove an important part of the dataset? Some degree of trustable control has to be maintained by the consumers of that data.

Operational arguments are similar. Unless the data is published via a service which can guarantee an SLA to the user, as opposed to the publisher, most businesses will shy away from using up-to-date Linked Data. Relying on a public endpoint is a business case that is more or less impossible to justify.

Paradoxically, the real value of that data is the ability to be downloaded in bulk. And that means that the difficulty, or ease, of using Linked Data is the same as that of using local RDF data. And thus uptake is only as easy as the easiest-to-use RDF store. Again, Dydra will help to lower the bar.

This personal emphasis on downloadable bulk data is why I'll be sad to see projects like data.gov go, as is currently suspected, and why every public Dydra repository provides a set of easy, obvious links for downloading the raw data, and not just a SPARQL endpoint. Data.gov provides easy-to-find dumps of each dataset in multiple formats, often including RDF. You don't need to depend on data.gov's continued existence to use the datasets there. It's the right model to build real applications. (Not to mention that there's some cool datasets there, too.)

Whether or not Dydra is the answer, we sure hope that RDF and Linked Data take off. If not, watching the world badly re-invent them for the next 20 years of my career is going to be frustrating indeed!

Do more with your shopping history

By Tara Hunt, aka MissRogue, CEO @ Buyosphere, who is a self-professed shopaholic and longtime online community member (about 18 years now). She wrote a book on this stuff and is completely nuts about bringing humanity back into business. She's worked with over 30 startups and been a serial entrepreneur.

A couple of years ago, I decided that I needed to get myself a good digital video camera for shooting the odd video podcast and interview. My budget wasn't huge, but I wanted to get something nicer for my money.

I searched in all of the usual ways: asking friends who I thought were knowledgeable, Googling for answers, going to the trusted brand sites. All roads left me even more confused. I gave up and walked into a photography store where the sales associate helped me pick out a decent camera. At that point, I was just ready to be done with the process, so I bought it on the spot. A year later, the FlipCam came out and I've used that camera way more than I ever used my fairly expensive video camera and it cost me $250 (the original purchase set me back $900).

This is what Barry Schwartz called the Paradox of Choice: when there are many choices but none of them are clear. The biggest paradox is that, no matter the choice we make, we are left wondering if we should have made a different one, leaving us ultimately unsatisfied with the results. It's also the main issue with shopping online. There is bountiful choice, but very little real guidance as to what would suit our individual needs.

Sure, there have been advances due to customer reviews and social recommendations ("people who bought this also bought this"), but by-and-large these just confuse us more. On any given product there will be a wide range of reviews without much context. A popular item with hundreds of reviews will generally show an equal number of one and five star reviews - all for a variety of reasons. A simple kettle purchase on Amazon.com can cause extreme angst. More information does not always lead to better information.

But it doesn't have to be this way. In fact, with the growth of the social web, more and more people are broadcasting who we are, what we want, where we go and even what we are eating. With every tweet, post and photo we are telling the world a bit about who we are. This data is still largely untapped. We spend all of this time crafting a great picture of ourselves and within days, even hours, it is buried and gone. Sure, there are a few marketing engines that are trying to mine our social media trails and push out contextual advertising, but these seem general and bland compared to how much thought, time and energy we put into signaling who we are and what we are about. We are no longer demographics or psychographics. There is something more individual happening that is yet to be explored.

Twitter and Facebook were just fledglings. Foursquare, Foodspotting and Fashism didn't even exist. But I had been blogging and leaving a digital trail of my own taste signals across the web since I started participating in forums and blogging. And I was definitely doing quite a bit of shopping online. If only I could somehow put this data together and take it with me everywhere I went to get better recommendations.

This is the issue that I started dreaming about solving about four years ago. Twitter and Facebook were just fledglings. Foursquare, Foodspotting and Fashism didn't even exist. But I had been blogging and leaving a digital trail of my own taste signals across the web since I started participating in forums and blogging. And I was definitely doing quite a bit of shopping online. If only I could somehow put this data together and take it with me everywhere I went to get better recommendations.

That is when I came across a project piloted by online marketing pioneer Doc Searls (co-author of The Cluetrain Manifesto) called Vendor Relationship Management (VRM). VRM is the reciprocal of Customer Relationship Management (CRM), the systems that businesses use to keep track of their customers. VRM's main thesis is if customers owned their own data and had tools to move it between vendors, customers would not only get better recommendations and have more control over their data, but vendors, too would benefit from the wider snapshot of a single customer's behavior. As soon as I heard Doc Searls talk about this, I knew it was the answer to my conundrum.

In regards to searching for the digital video camera, imagine this scenario: I indicate that I'm looking for a camera and agree to temporarily authenticate a checklist of information to any store that carries video cameras so that they can make better recommendations. I let them see my overall history and savvy with technology, the general price range I pay for equipment, my network of friends, my lifestyle and tastes and perhaps even things like my travel frequency (to understand how portable the camera should be). The stores would then send me back recommendations based on my preferences and show, if available, reviews and indications of anyone in my network that bought the recommended items. The retailers would have a better chance of making the sale and a stronger indication that I'm ready to make a purchase, saving them some serious marketing dollars. They could even pass the savings along to me. I would narrow my search and make it less confusing. Everyone wins.

That sounds great, doesn't it? Customers in control. Retailers passing marketing budgets along to savings for their customers. More certainty and efficiency with online shopping and decision making. Reduction in overall consumer anxiety. But it's still a long way off.

First off, customers don't own our own data. The closest we come is self-hosted blogs. We can't even search tweets back beyond a couple of months and there is currently no search within our own content in Facebook. Exporting raw data from any of

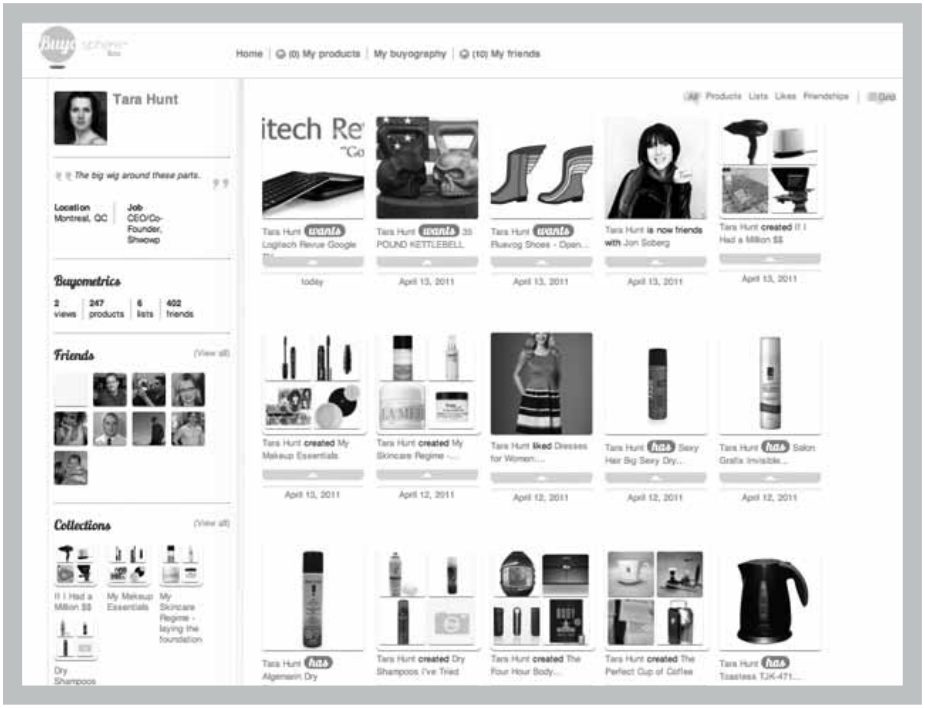
these networks would be basically useless because it's in a format that isn't consumable by any online retailers. It lacks context and the social ties that make it most interesting. I've also yet to find an 'export purchase history' button on any of my online accounts.

Secondly, retailers aren't set up to receive packets of data from individuals. Not are they only unable to receive the data, but most of the recommendation software they use still uses outdated demographic and psychographic slicing (think, "People who bought this also bought this"). Many of them are trying to set up their own networks that connect to Facebook or Twitter, which is not really a solution, either, because it just perpetuates the building of lost data into the current silos.

Luckily, there are a few projects emerging that address the first piece of the puzzle. One is my own startup, Buyosphere, where people can gather what they own and buy easily through various means. We are focusing on very specific data: the products we buy. There is also an open source project called Sing.ly (http://www.readwriteweb.com/archives/creator_of_instant_messaging_protocol_to_launch_ap.php) that aims to create a way for people to simply synch their data from any social network into a single self-hosted location. Once the data is gathered, the user can authenticate anyone to access it, making it portable. There is a growing group of companies emerging and collaborating at the Personal Data Ecosystem project (http://personaldataecosystem.org/) led by Kaliya Hamlin.

The hope is that the second piece - the ability for retailers to receive, parse and recommend with a better understanding of individual needs - will be built off of this more portable data as this market emerges.

The future makes much more sense if data is democratized, especially consumer data. It takes full advantage of the benefits of the social web, creates a model for understanding recommendations at the highly personal and individual level, and, ultimately, helps customers regain control of the data we are producing daily.



Linking Open Data at the University of Southampton

By Christopher Gutteridge, University of Southampton Linked Open Data Architect

I didn't get into Linked Data by quite the same route as most of the people I work with. I'm not a researcher - I'm a systems-programmer - and my experience is mostly programming database and web systems. My projects don't end when the funding runs out; they have to keep working as part of the infrastructure, year after year.

I'm web projects manager for The School of Electronics and Computer Science (ECS) at the University of Southampton. ECS has more than its fair share of professors who are big names in the web, such as Sir Tim Berners-Lee, Dame Wendy Hall and Nigel Shadbolt. Nigel advises the UK government on Linked Open Data, so it was a no brainer that http://data.southampton.ac.uk/ would soon incarnate in some fashion, but the devil is in the details.

A few years back, Dr Nick Gibbins ran a project to create RDF open data for our school, http://id.ecs.soton.ac.uk/docs/ and from this I learned lots about good URI design but my RDF skills were still limited, so I spent much of 2010 brushing up my skills by building practice datasets, on http://data.totl.net/. One of the things I've learned from being a sysprog (and from making pancakes) is that your first attempt is always flawed in some way, so it's worth putting in the hours to get over some of the obvious beginners mistakes before you start working on something that really matters.

Much discussion was being had above my pay grade. This resulted in some very high level backing and the scariest meeting of my life. In addition to the academics (Nigel & Wendy, and Hugh Davis our head of e-Learning) there were some people who I didn't really know at the time, but were key to the success of the project. They were the heads of the University IT, communications and finance departments. Working with their authority opened up many doors which would normally have been closed to a research project. Or to put it another way, dropping their names cut a hell of a lot of red tape – it felt like having a letter with the King's Stamp, reading "Let this man have the data he requests without let or hindrance".

Lots of ideas were discussed but the broad plan was that we should work on a very tight timescale and produce a public beta within 4 months. My remit was: 'anything not confidential', but leave 'data produced as a result of research' as a task for later. A rule of thumb being: if it was subject to a FOI (Freedom of Information) request, we might as well publish it. It was also agreed that any tools I produced as a result of this work would be open sourced. *https://github.com/cgutteridge/Grinder*

I set myself a couple of additional guidelines too. I didn't want this to be exploration – we've had plenty of that in Linked/Semantic/Open data. What I wanted to do was trailblazing – using tools and techniques that could be reliably deployed at other organisations.

Another issue was that I needed to win over the data holders. Experience has taught me that administrative staff don't appreciate being given new responsibilities as they tend to already have a full (or over-full) dance-card. So I tried to start by finding out what they already did, what data they had and what issues they had. The most successful example is catering.

Catering already maintained all the data I wanted to publish as they needed for their own operations; planning, printing menus, making websites. Much of it was in a spreadsheet, but with lots of little niggles like inconsistent style and hard-to-parse layout. I was nervous, but they were very receptive to my suggestion of my rewriting their menus and spreadsheets. My revised spreadsheets are no harder for them to maintain but are much easier to parse unambiguously. To make things even easier, we shifted from Excel to Google Docs. This let me and them edit the same document, and easily exported to CSV for automating the process. Using colour and other styles make looking after the spreadsheets easier for humans.

The Raw data is available as CSV from; *http://data.southampton.ac.uk/dataset/catering.html*

The final result for catering is that they are not doing significantly more work than they were previously. Now all of their points of service have a nice webpage including

opening hours and price lists, so they don't need to fuss about maintaining their own pages. Soon, we hope to be able to reprint individual menus for each point of service when prices change without having to fiddle. http://data.southampton.ac.uk/point-of-service/38-lattes.html?view=menu

We're also looking into cute little things like mobile menus

http://data.southampton.ac.uk/point-of-service/38-lattes.html?view=mob

But the key message here is quid-pro-quo. Admin staff don't have the luxury of doing things out of idealism. They don't have the slack to care about Open data (let alone Linked Data) unless it makes their life easier or provides an obvious benefit for the users of that service. What we've achieved isn't perfection; it's a sustainable service on time and on budget.

Off the Shelf

There's not that much software, yet, which produces RDF out-of-the-box, but we use two at Southampton: Agresso financial software and EPrints respository software (which I wrote the RDF extension for). The RDF for both of these projects are work in progress, but it makes more sense to feed back suggestions and requirements to EPrints and Agresso rather than try and implement our own solution.

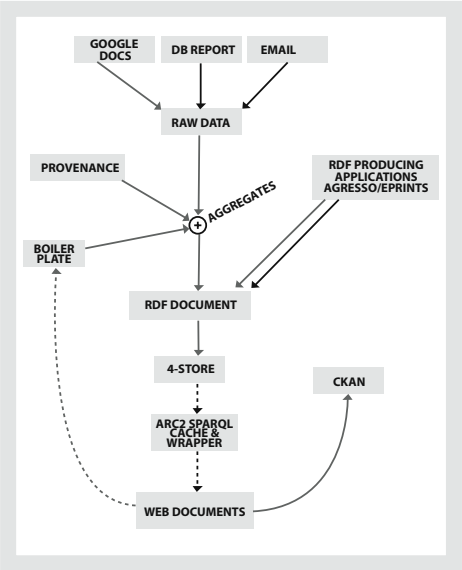
Minting URIs

A real challenge is minting URIs when the data changes and the people maintaining the data don't have years of experience in information theory. When the catering department says "Sandwich – Price Band B" they don't actually mean that at all. Fruit Salads are sold as "Sandwich – Price Band B".

Even things with more solid looking identifiers, like rooms, are imperfect because our university reuses building numbers from old buildings (after a suitable period of mourning; a decade or more). There is no semantic scheme to uniquely identify a building throughout all time and space. To address this situation I have ignore the problem. It's an imperfect world and confusions like this are the 404s of the next web. They suck but we can't let the risk paralyse us.

Don't diss 404s, they are what make a billion-page web possible. Perfection doesn't scale and it's the ones with the cracks which let the light in.

We use id.southampton.ac.uk for all URIs and data.southampton.ac.uk for HTML and RDF documents. It's an unusual pattern but much clearer, to people new to Linked Data, that id.southampton serves a different purpose to data.southampton.



Namespaces for predicates and classes

To get the data right, where possible I've collaborated with peers at Southampton and other Universities. I've asked on mailing lists. I've done many web-searches. However you get to the point where you've just got to bite the bullet and make a choice and I've done that with the data structures on data.southampton.ac.uk – I'm sure some will prove to be mistakes but the alternative is paralysis.

Where possible I've used existing classes and predicates – dcterms, foaf, skos, etc. but often I've had to mint my own. These fall into two categories.

http://id.southampton.ac.uk/ns/

Classes & Predicates which I really don't think anybody else should re-use. Either I know there's someone designing an ontology, or I wish there was, or they are just entirely specific to the university or our systems and in no way is it advisable to reuse them. In these cases we mint them at http://id.southampton.ac.uk/ns/ . We have not got a schema up at that URL, rather we try and make very long and clear textual descriptions in the URI so that confusion will be minimal.

http://purl.org/openorg/

Every now and then I find a class or predicate that really should exist but does not seem to in any common namespace. In these cases I've defined them, with advice and suggestions from peers, under http://purl.org/openorg/ . This namespace is intended to augment existing ontologies for organisations wanting to provide open organisational data. The base of a namespace shouldn't be political, but it's my belief that other organisations will be more comforatable with a non partisan namespace. Other organizations are welcome to suggest additions to this namespace, with me as benevolent dictator.

Linked Open Data 2.0

I'm just being deliberately annoying with the title, but I really do think that we've entering a new phase with linked data. Up until now (with the possible exception of FOAF profiles) most open data is exposing specific collections of "special" data. What's about to happen is a profusion of data sites which contain a bunch of datasets with heaps of overlap. The structure of administrative data about ten-thousand organisations isn't ten-thousand unique snowflakes, it's ten-thousand things which are, more or less, identical (like snowflakes).

What we need to be doing is laying down sensible semantic patterns for recurring datasets: People, Places, Products, Points-of-Service, Prices, Power usage, Pedagogy, Publications, Pubs, Payments and Events. (My thesaurus has no synonyms for events which begin with "P")

Not every university, or indeed organisation, needs to design an ontology, or even how to use it. We need to start providing the community with sensible, semantically viable, RDF structures to just copy. The HTML web was built on copying. To try and nudge this forward we've set up openorg.ecs.soton.ac.uk to try and collect sensible patterns for every-day RDF datasets. For a pattern to be considered mature it must be used by multiple unconnected organisations and have tools to both produce and consume it.

I know, as programmers, we can see the value of producing data in machine-readable form, but most people don't see things in that abstract way. We need consuming tools to bribe web-managers with. "If you produce data in this way, you can do all this cool stuff with it right away". The next wave, after early adopters, will only be able to justify providing Linked Open Data as soon as there's some perceived value to them, so they can justify the costs. Cool little web-embeddable gizmos may help with that. Really useful and powerful RSS tools didn't come along until RSS was widely used but RSS didn't become widely used until there were at least a few really useful tools. Things like Good Relations are a great start as they publish worked examples along with the ontology, but better still would be tools to validate and render the data.

Climbing the Learning Curve: A US Government Milestone for Publishing Linked Open Data

By Bernadette Hyland, CEO, Talis Inc.

Starting in 2007, early adopters of Linked Data publishing suggested there were dramatic cost savings in data sharing, maintenance and application development. While these innovators acknowledged the International Data Standards for the Web were solid, the tools for developers were a potpourri of Unix scripts written by people adept with the nuances of semantic technologies. Early enterprise Linked Data projects were created by innovative teams, and several enjoyed the support of academic research groups with specialized skill-sets.

Linked Data Changes the Status Quo

However, just four years later in 2011, we are seeing cities, states and federal governments using Linked Open Data to help change the status quo. In the past, governments carefully collected and curated databases of information. Any member of the public who wanted to know would have to make a formal request so a government employee could retrieve the requested information. In the US, this process is the FCC Freedom of Information Act also called FOIA.

With Linked Open Data, the contents of those databases can be published directly to the Web and citizens can directly access the information. That saves time and effort on the government's part responding to these requests. Having to do more with less, we are witnessing firsthand how government agencies are working to reduce inefficiencies and transform how they disseminate information internally, across agencies, and with the public.

In tight economic conditions, identifying new insights from increasingly available data provides citizens, policy makers and administrators access to actionable information. This approach has immediate benefits to millions of people facing the aftermath of natural and man-made disasters, crop failures, global warming, etc.

Like many governments around the world, the US Administration's Open Government Initiative has been defined in a tight fiscal climate. Like other executive administrators, the Obama Administration is supportive of more modern approaches to government transparency and efficiency. A guiding principal of Open Government projects is to institutionalize making data more widely available in usable formats via the Web. Making data available via RDF (Resource Description Framework) represents a dramatic and sustainable cost reduction for data sharing, maintenance and application development.

Talis recently assisted the US Environmental Protection Agency's Office of Environmental Information to model and publish an important data set called the Facility Registry System (FRS) as high quality Linked Open Data. FRS lists over 2.6 million facilities in the United States that are of interest to the Environmental Protection Agency (EPA) for the purposes of environmental monitoring. The FRS represents an aggregation of data, integrated and validated from over thirty federal databases and over fifty US State, Territory and Tribal databases.

In tight economic conditions, identifying new insights from increasingly available data provides citizens, policy makers and administrators access to actionable information.

The EPA brought together Linked Data experts from Talis, along with FRS subject matter experts from EPA and their contractors. We held face-to-face sessions during February and March 2011 to walk through the data modelling process, validate the data and look at various visualizations on the newly created Linked Open Data set. Data modelling involved a two day deep dive on the existing relational database model, followed by an interim model review.

The team discussed strategies for URI selection, vocabulary re-use, creation, and query support via SPARQL. Talis delivered a set of sample SPARQL queries, a comprehensive modelling guide, a CIO Management Briefing, conversion scripts, and the RDF for 2.6M Facilities represented in approximately 103M RDF triples.

Together with EPA, we identified next steps and reviewed future maintenance activities. US EPA now has a 'reference implementation' for a high quality data set modelled and converted to RDF that can be used for future Linked Open Data initiatives. Armed with this information, we believe the EPA team is feeling prepared to support the ongoing maintenance of the FRS and future Linked Data sets.

Another key deliverable was to present why senior management should support future projects, especially since they are already publishing information on their Agency's web site. Simply put, the ease and speed of merging databases via RDF, overcomes an array of issues associated with data silos. This comes as no surprise to those of us who are proponents of Linked Data however; the direct experience of these benefits by the customer was an important milestone.

The Fine Print

Let's be realistic: Linked Open Data is not a silver bullet. There is a learning curve associated with employing a new approach to publishing information. This includes:

1. Thinking of data in terms of how it is related to other things;
2. Documenting data in both HTML and machine readable formats such that it can stand on its own; and
3. Recognizing the social contract you are making by publishing information on the Linked Data cloud.

Unlike in the traditional 3-tier application world where an enterprise developer can walk down the hall and visit a database administrator, in the world of Web accessible data, useful data must be 'self-describing.' As respected ontologists Dr. Dean Allemang and Professor Jim Hendler point out in their book Semantic Web for the Working Ontologist, "the advantages

of how RDF works is that it is possible to query a dataset without knowing anything about the data set at the outset."

Linked Open Data (LOD) actually uses many traditional data modelling principals, with the following differences. An LOD data modeller:

1. Makes explicit the implicit relationships in the data;
2. Expresses those relationships using URIs;
3. Re-uses (authoritative) vocabularies wherever possible; and
4. Publishes data as RDF (an international data representation standard) to facilitate data merging.

Publishers Make Social Contracts

Publishing LOD involves management commitment. As a Linked Open Data publisher, there is an implicit commitment to publicly maintain it. This can lead to overhead of publishing updates which must be performed either by agency staff or a government contractor. There are two main compromise routes to address this. Either, publish on the understanding that there will not be a change until the next version, whilst providing a simple route for people to report errors and request changes that may appear in that next version.

Alternatively, provide a near live view of constantly evolving data that will reflect such requests as part of normal additions. This second approach may well not be applicable to data with necessary regular publishing schedules, such as annual reports. Your agency will need to commit internally to replication of database dumps on a routine basis and maintenance of data conversion scripts (e.g. to reflect changes in a database schema).

As a Linked Data publisher, you are making a social contract with future users you are not likely to ever know or meet. This warrants thoughtful, up-front planning with Linked Data experts. For example, if an authoritative government agency publishes information that people in turn begin to integrate, due care and consideration must be given to formulating URI patterns *and maintaining them*.

Move Forward Incrementally

LOD modelling does require some specific skills, but it isn't rocket science. Those familiar with traditional data modelling and governance issues can come up to speed with reasonable attention to the differences between tradition and Linked Data modelling.

Model one useful data set and visualize it with complementary data sets. On this project we used the EPA FRS data set, plus Geonames and Tom Heath's Near data. Within a day we quickly joined data sets and created compelling visualizations -- proof of the Linked Data approach! Compare our 6-8 week part time modelling and development cycle to a typical enterprise database migration or addition to a master data management project. Experts agree, it usually requires six or more months using the relational approach, before the application development even begins. This is not meant to imply relational databases are going away or are bad. That simply is not true. Relational databases are the right tool for many enterprise data management requirements, especially secure transactional based solutions.

However, for publishing many forms of government collected information, exposing the contents of a relational database as Linked Data is a straightforward process starting with data modelling, publication of flat files via the World Wide Web, that can be queried using another International Standard called SPARQL. Note, there is one SPARQL standard that either vendors support or not.

During this EPA FRS project, the benefits of an LOD approach were derived within two months, inclusive of several days of face to face data modelling and presentation sessions, development of conversion scripts (scripts to convert CVS files to RDF), and a series of SPARQL queries against mostly Open Source visualization tools. Guidance from subject matter experts to assist with initial data modelling activities, in short, defined sprints is recommended. Depending upon the complexity of the data set, domain matter experts add tremendous value. A question we are often asked is:

"what is the value proposition of Linked Open Data? Why should we care?"

The value proposition of Linked Open Data is in the international data standard format RDF. There are multiple serializations of RDF, but it is trivial to go from one to another.

Why is RDF so Important?

The RDF Data Standard, guided by the World Wide Web Consortium (W3C) took many years of the world's leading research scientists, academics and software vendors to drive toward consensus. Publishing and consuming RDF leads to greater efficiencies in terms of data sharing, reduced costs for application development and ad hoc visualizations. The use of RDF is the foundation for the significant reduced costs for both publishers and consumers of data, in the public and private sectors. In turn, the broad use of Linked Data by leading retailers, manufacturers, compliance, legal and financial services organizations is driving a global economic and political transformation.

Linked Data infrastructure based on RDF allows organizations to Easily exchange, re-use and remix content.

Thus, with an ever-increasing number of Open Source tools, one can rapidly convert, publish, visualize and share data. The question becomes, can you afford not to leverage Linked Open Data?

Importantly, spurring economic development through new businesses starting or expanding with improved access to information is a cornerstone of the current US Administration. Reputable commercial companies, some of whom are based on large Open Source Projects, are helping government agencies handle data modelling, user interface and data driven application development. It is recommended that any organization using Open Source or proprietary solutions have an appropriate support agreement in place with their vendors. The use of Open Source Software does not eliminate the requirement for a service level agreement with a vendor if this application is in production.

"Third parties, in turn, are consuming this data to build new businesses, streamline online commerce, accelerate scientific progress, and enhance the democratic process.[1]
-Tom Heath, Christian Bizer

In the last several years we've seen a marked shift in government from a "need to know" to a "need to share" mentality. We are seeing government agencies define guidance and best practices for use of Facebook, Twitter, blogs, podcasts, RSS feeds, YouTube, and most recently, data driven applications based on Linked Open Data (LOD). Publishing Linked Open Data is increasingly seen as a serious alternative approach for faster, smarter, cheaper data integration projects.

"Cooperation without coordination is what makes RDF so powerful."[2]
- David Wood

One of the major insights gained upon completion of the Facilities Linked Data project was how easy it was to put the data into any standards compliant RDF database. There were no complex conversion scripts to load the data into different databases. Using International

Data Standards, it is trivial to load RDF into a store, query and visualize the data in literally one day. This same effort using a traditional 3 tier approach using relational technologies would have taken six months or longer. This is a major benefit and cost savings to federal systems integrators and Government Agencies, resulting in significant cost savings and time efficiencies.

Another major insight for the project participants was the ability able to view their data through a variety of freely available visualization tools. Several of the tools demonstrated were developed by or with support from Talis including LinkSailor, Talis Platform, Callimachus and Morph. Two other useful visualization tools demonstrated were Spark and Exhibit - both available as Open Source.

The first thing I looked at was, "what facilities of interest are in my backyard?" We all care about our communities. Using the EPA data set, I was quickly able to see exactly where facilities of interest were relative to my home and my children's school, both of which were of interest to me.

With more and more government agencies publishing data we can readily understand and visualize via the Web, the more citizens are able to make informed decisions. The easier it is for government authorities and public industry to access and visualize high quality data published by authoritative sources such as the US EPA, the more efficiently they can manage their limited resources, and do so responsibly.

With projects such as the EPA's Facilities Registry published as high quality Linked Open Data, everyone wins. The EPA, other government agencies, citizens, journalists and bloggers will all be able view valuable data and create mash ups with well-modelled Linked Open Data sets. We look forward to the EPA Facilities data set being made available via data.gov in the near future.

How Can We Get Started?

Books in hardcopy or electronic copy are available, including Linked Data: Evolving the Web into a Global Data Space by Tom Heath, Christian Bizer, (publisher: Morgan & Claypool 2011), and Linking Enterprise Data, David Wood, editor, (publisher Springer 2010). Guidance from a knowledgeable, helpful global community via the LOD mailing list and related web sites is a great resource.

The shared collective knowledge, combined with mature Web Architecture Standards published by the World Wide Web Consortium (W3C) should provide comfort to senior managers that a Linked Open Data approach is not bleeding edge nor a fad. Instead, LOD and data driven applications, ideally via mobile devices, are the future.

International Data Standards, combined with flexible developer frameworks, many of which are available as Open Source, are now enabling major high profile projects. These are found today within public bodies in the USA, UK, Australia, Denmark and many other federal governments. Whether you are a Web-native company such Google, Yahoo! or Amazon, or a venerable newspaper such as The New York Times or The Guardian, or a publisher such as BBC and Wolters Kluwer, or world-class medical centre such as the Cleveland Clinic, you can benefit from faster, smarter, cheaper data integration leveraging Linked Open Data.

For Linked Open Data to become the de facto approach for data transparency, we need published reports with defined costs and return on investment (ROI) figures. We encourage you to write and share your experiences in relation to producing and using Linked Open Data.

Participation in the W3C Government Linked Data Working Group

If you are interested in participating in the process of standards creation to help governments around the world publish their data as useable Linked Data, we encourage you to look into the W3C Government Linked Data Working Group. This group is exclusively focused on data publication using Semantic Web standards, deployed on the Web following linked data principles, as introduced by Tim Berners-Lee in 2006, in Linked Data.

Conclusion

Our experience working with the US EPA Office of Environmental Information suggests that agencies are going well beyond publishing recommendations. Committed advocates of Open Government are actively producing the guidance for procurement, vocabulary selection, URI selection and management, best practices, and identifying the licensing and authority issues. They are publishing Linked Data and making it available via catalogue sites for their respective government agency.

It would be hard to argue the economic and political benefits that have been realized through the advances in the last twenty years of the World Wide Web. However for Linked Open Data to become the de facto approach for data transparency, we need published reports with defined costs and return on investment (ROI) figures. We encourage you to write and share your experiences in relation to producing and using Linked Open Data.

With high quality data sets published on the Web via sites such as data.gov, data.gov.uk and many others, data driven applications will increase citizens' understanding of governments and provide a mechanism for governments to be more responsive to their constituents.

Linked Open Data is a key component of a well-informed society. Few would argue that well-informed citizens are the foundation of functional societies and a balanced global economy. We hope that you will join the data ecosystem that is driving transformation around the globe. Become a publisher and/or consumer of Linked Open Data today!

References:

[1] Linked Data: Evolving the Web into a Global Data Space by Tom Heath, Christian Bizer, publishers: Morgan & Claypool (2011).

[2] The Semantic Web enables cooperation without coordination." – David Wood, Semantic Web Elevator Pitch, see http://semanticweb.com/semantic-web-elevator-pitch-for-journalistic-research_b17662

Sources:

1. Early adopters of Linked Data, http://3roundstones.com/led_book/led-hyland.html
2. International Data Standards of the Web, http://www.w3.org/standards/semanticweb/data
3. FCC Freedom of Information Act, http://www.fcc.gov/foia/
4. US Open Government Initiative, http://www.whitehouse.gov/open
5. US EPA Facility Registry System, http://iaspub.epa.gov/sor_internet/registry/facilreg/home/overview/home.do
6. Tom Heath's Near Data, http://blogs.talis.com/research/2011/04/14/follow-your-nose-across-the-globe/
7. Tom Heath and Christian Bizer, Linked Data: Evolving the Web into a Global Data Space, Synthesis Lectures on The Semantic Web: Theory and Technology #1, Morgan & Claypool Publishers 2011
8. The Semantic Web enables cooperation without coordination." – David Wood, Semantic Web Elevator Pitch, see http://semanticweb.com/semantic-web-elevator-pitch-for-journalistic-research_b17662
9. Dean Allemang and Jim Hendler, Semantic Web for the Working Ontologist: Effective Modeling in RDF and OWL. Morgan Kaufmann, 2008
10. http://www.w3.org/RDF/
11. Link Sailor, http://linksailor.com/nav
12. Talis Platform, http://www.talis.com/platform/
13. Callimachus for Semantic Application Development, http://sites.google.com/site/callimachusproject/
14. Morph Semantic Web Formats Converter, http://morph.talis.com/
15. Spark for Mashups, http://km.aifb.kit.edu/sites/spark/
16. Exhibit for Faceted Navigation, http://simile-widgets.org/exhibit/
17. W3C Government Linked Data Working Group, http://www.w3.org/2011/gld/
18. W3C Linked Open Data Mailing List, http://lists.w3.org/Archives/Public/public-lod/

'Follow Your Nose' Across the Globe

By Tom Heath, Lead Researcher, Talis

Abstract: The ability to follow links from one data source to another is just one way in which Linked Data excels over any other method for publishing data on the Web. These links are the very fabric that make up the Web of Data, enabling ad-hoc discovery of new information by simply following the connections between data records. We call this principle follow-your-nose. When we walk or drive around unfamiliar locations we often use the same principle to find our way – we follow our noses to our intended destination. In this article I'll describe how the Talis Research team have begun to interlink geographical points of interest from a number of different Linked Data sets, in the process creating a new data set that enables you to follow your nose across the globe.

Imagine you've just arrived in an unfamiliar place, perhaps on a business trip (or recently beamed down from the Starship Enterprise). One of the first things you'll probably want to do is find out what things are nearby. Google Maps provides a great "search nearby" function – try searching for a specific location, then search nearby, and enter just a * to get everything. However, this functionality is geared heavily towards business listings, and the data isn't exactly open, making it hard to reuse in other applications. We wanted to try something similar, using the growing range of liberally-licensed Linked Data [http://linkeddata.org/] sets with a geographic component. Here's what we did...

We started with the Geonames *[http://www.geonames.org/]* data set, all the geo-coded entries from DBpedia *[http://dbpedia.org/],* and data about schools from data.gov.uk *[http://data.gov.uk/]* -- we'll call all the things described in these data sets points of interest (POIs). In truth many of the entries in Geonames and DBpedia are not points but regions or areas, but that's an issue for another day.

Next, we implemented some cheap and cheerful secret sauce to identify, for each point of interest, all the other POIs that are 'nearby'. For each of these pairwise relationships we created two new RDF triples, one stating 'x near y', the other 'y near x'. For example, our algorithm may reveal that Birkby Junior School [http://education.data.gov.uk/id/school/107626] in Huddersfield, UK, is near to Kirklees Incinerator [http://dbpedia.org/resource/Kirklees_Incinerator], in which case we simply create the following RDF triples:

```
<http://education.data.gov.uk/id/school/107626>
    <http://open.vocab.org/terms/near>
    <http://dbpedia.org/resource/Kirklees_Incinerator> .
<http://dbpedia.org/resource/Kirklees_Incinerator>
    <http://open.vocab.org/terms/near>
    <http://education.data.gov.uk/id/school/107626> .
```

We do the same for all pair-wise relations in each cluster of 'nearby' items.

The process runs on top of the Hadoop framework [http://hadoop.apache.org/] for crunching large volumes of data, and produces a new data set of more than 700 million RDF triples that represents a significant increase in interconnectivity within and between the input data sets. Along the way we also create URIs for all the intersection points in the lat/long grid (down to a specific level of granularity). These are linked to the URIs of nearby things from the input data sets, and to each other.

Our definition of near was deliberately left loosely defined. As a concept it is far too subjective and context dependent to try and define more precisely, so we kept it vague and left plenty of scope for refinement according to the needs of specific applications. We achieve this in the following ways:

Imagine you've just arrived in an unfamiliar place, perhaps on a business trip (or recently beamed down from the Starship Enterprise). One of the first things you'll probably want to do is find out what things are nearby.

Firstly, imagine an application uses this data set to find all the other things near to a particular point of interest, but also wants to know the distances to each and which is the closest. The data to answer these questions doesn't need to be materialised in the data set. Instead, the consuming application can lookup the URIs of the nearby points of interest, retrieve the data describing the POIs to obtain their original geo-coordinates, and perform the post-processing required to compute the distances between points and find which is nearest.

Extending this principle, it's also trivial to use this data as input to a matching process that identifies when the same entity is described in different data sets (with each assigning a different URI for the item). For example, our approach finds that the item identified by the URI <http://dbpedia.org/resource/Gad%27s_Hill_School> is near to a number of other things, including the item identified by the URI <http://education.data.gov.uk/id/school/118944>. In fact, both these URIs identify the same thing, Gad's Hill School, and could be connected using a sameAs relationship. The cost of assessing whether these two URIs identify the same thing, perhaps by computing the string similarity of the labels assigned to them in each data set, is much lower when using the set of near links as input compared to using both data sets in their entirety, as the number of pairwise comparisons that must be made is significantly reduced.

The second mechanism we introduced to ensure the data set could be reused and refined where required was to make the data set as navigable as possible. A typical Web API requires each query to be formulated afresh before it is sent, meaning the data set isn't easily browsed in an ad-hoc fashion. Instead, logic is required at the client/application-side in order to formulate the next query, before the client or user can move from one record to another. In contrast, this new data set applies – to the physical world – the same follow your nose concept that's so central to Linked Data. The concept is blindingly obvious when you stop and think about it. We navigate the physical world by following our noses, looking for landmarks, finding our bearings, and adjusting our courses – why should it be any different when we navigate the Web from a

geographical perspective?

To support this mechanism, the data set includes links that connect each intersection point in the latitude/longitude grid to those that surround it. These links are expressed in terms of compass bearings, allowing an application to move from grid point to grid point in whatever direction they choose, potentially traversing the globe in the process. In reality there are some gaps in coverage, primarily corresponding to oceans and areas of low population density that are typically under-represented in the input data sets; as we currently only materialise data for intersection points that have nearby POIs, these areas are generally not covered by the grid.

We have a number of plans for enhancements to the data set, but in the meantime we've made an initial release of the data set, which is hosted in a Talis Platform store and exposed through the API at *http://rdfize.com/geo/point/ [http://rdfize.com/geo/point/]*. The goal of this API is to provide a simple access mechanism for application developers wishing to find all points of interest near to a particular location. To achieve this, developers simply need to construct URIs of the form *http://rdfize.com/geo/point/latitude/longitude/* e.g. *http://rdfize.com/geo/point/9.022762/38.746719/ [http://rdfize.com/geo/point/9.022762/38.746719/]* (a point close to the centre of Addis Ababa).

> *As a concept it is far too subjective and context dependent to try and define more precisely, so we kept it vague and left plenty of scope for refinement according to the needs of specific applications.*

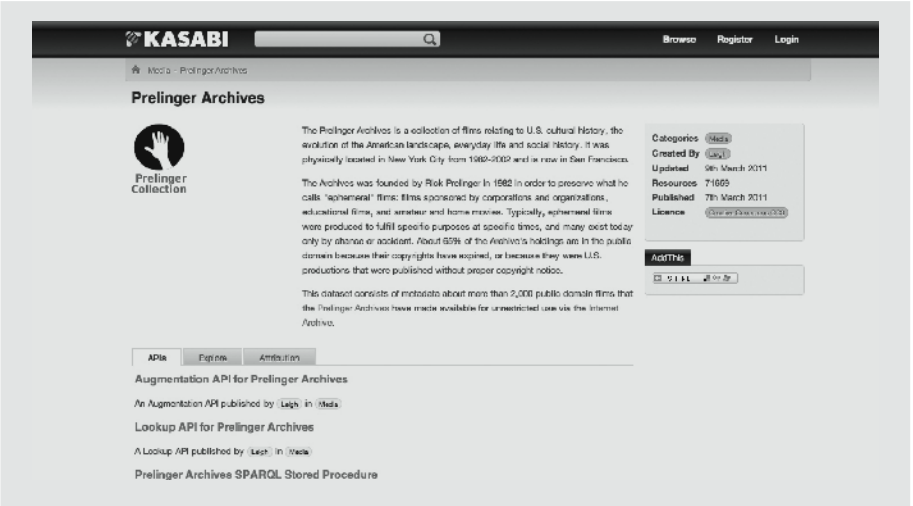
Clients or applications that look up these URIs will be redirected to a URI that identifies the nearest grid intersection point (in this case http://rdfize.com/geo/point/9.02/38.75/ *[http://rdfize.com/geo/point/9.02/38.75/].* From there the client will be redirected to an HTML or RDF description of that point. Currently we support the RDF/XML, Turtle, N-Triples and RDF/JSON serialisations of RDF, as well as a simple tabular HTML view of the data for human users.

Those looking for a more polished interaction with the data set should sign up for the beta of Kasabi [http://kasabi.com/], where we plan to release enhanced versions of the data set in the coming months.

Unveiling Kasabi

By Zach Beauvais

For the past few months, I have had the opportunity to work on something new at Talis: something that brings together big datasets, interesting concepts and developers in a single, rather exciting project. You will have worked out from the title that it's called Kasabi, but hopefully I can still surprise you with what we're building it to do.



What's a Kasabi?

For the bullet-point enthusiast, Kasabi is:

- A place to expose data to a wide audience
- A place to browse and use datasets
- A place to fuel innovation and business
- In Beta

Think of Kasabi as an environment where people with data can make it available to people who want or need it. You can think of it like a marketplace for data: providers can advertise their wares and sell their valuable products to throngs of eager developers and other interested customers. On the other hand, developers, information architects, start-ups or just interested people can browse along the fronts of vendors' stalls before finding a batch that matches their requirements. We're working on features currently which will let vendors set prices and tweak terms and conditions, and have all sorts of ways for interested people to find interesting datasets.

But the marketplace metaphor only gets us part of the way. In an actual market, a vendor can't rapidly tweak his business model or explore the potential of a product by changing the way it's put together, or put more work into part of a product and let another tick-over because customers need something slightly different. Nor can an interested shopper find something that is "kind of like what I want, but in the wrong shape," and alter it there and then in the market. Nor, once they make these changes can they tell all the other vendors and shoppers how they changed it to improve it for themselves.

So, the market is helpful for illustrating what Kasabi basically is, but doesn't really paint a complete picture—and I am in danger of over stretching it.

You're the tailor.

Most web applications have an API: a way to access a selection of data they provide. This lets developers get at specific information from a service and build on it, or tie it in with their own project. Kasabi brings this application-style approach to hosted datasets. In essence, data providers can publish their sets with APIs that let developers access them according to their needs. So there isn't a single over-arching Kasabi API, but rather APIs for datasets. Also, Kasabi hosts datasets with a range of different APIs for each set. So every set is given its own suite of APIs providing different views into the data.

As a customer, I could browse the catalogue of datasets, or search for a particular kind of data and have a good look at it on the shelf (as it were). I can find out what kind of data it contains, who has made it and when, and even read some quite detailed information on how it was put together. I then have a range of different ways I can access it. If one of its APIs suits my needs, I can select it, make my payment (if any) and begin using the data in my project.

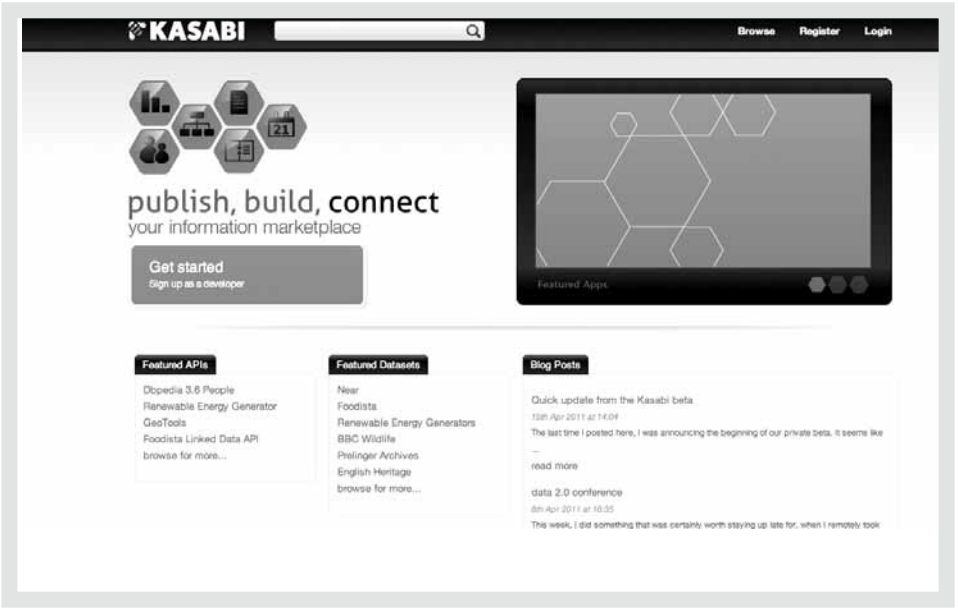
Lets say, however, that I can find a dataset that contains the kind of information I need for a project, but none of the APIs fits my needs. In Kasabi, I can actually design my own API for the dataset. I can then choose the best kind of API, change the output format, and tweak parameters until I've got a contextual view of the data that suits me better.

Once I have my tailored API, I can also share my version with the rest of the Kasabi community. So datasets in Kasabi have a suite of core APIs off-the-shelf, and the possibility of user-contributed APIs to further augment access to each set. This not only lets developers finely-tune the data-access they need, but it gives providers effective crowd-source API development.

Data in a graph.

Kasabi stores data as RDF triples, giving us a great foundation for managing the different kinds of data we host. The graph structure of the data underpins our offering of APIs, and has given us a good framework for the data publishing curation features being created for data providers. Use of RDF means that data we host is also available as Linked Data, providing yet another way to access the data. You can browse the Linked Data structure of datasets, and have access to the powerful SPARQL query language for every set. SPARQL also forms the basis of some of the APIs in Kasabi, and it's proven a natural tool for tailored access to data.

However, this does not mean that every developer in Kasabi needs to be au fait with RDF. It doesn't mean that you need to have earned your flight wings in SPARQL just to get at the data. Datasets can be accessed in more familiar ways (i.e. REST) and through more innovative tools. It's part of our vision to make access to data easy and powerful. We use Linked Data and the Talis Platform for the potential of fine-slicing of datasets, graph patten matching, and the web-native publishing of RDF and Links. But we've also made it possible



to access this information through reliable technologies developers across the web are familiar with.

For more information on the Linked Data side of Kasabi, I can recommend a piece written by Kasabi's programme manager, Leigh Dodds, at: bit.ly/contextking.

What's in the Kasabi Beta?

Kasabi is currently in beta, and depending on when you read this article, it will either be in its early, invite-only form, or in its more public-access stage. In either case, you can find out more about it at kasabi.com, where you can sign up to take part. This tricky timing also means that anything I describe in the beta here may have been expanded or tweaked, so I can only describe what is in the beta now, and invite you to come visit to find out where we're at when you read this.

It's been a few hundred words without bullet-points now, so I'll add in a few more for the skimmers. In a nutshell, the Kasabi Beta comprises:

- select pre-loaded datasets (some new!)
- discovery tools: search, browse, explore
- range of APIs for each dataset
- API tailoring
- share and document SPARQL queries
- initial client libraries in Ruby and PHP

Developers currently have access to a catalogue of pre-selected datasets hosted in the Kasabi environment. Some of these sets are new, others may be more familiar, but all are available to any beta tester. Some of the sets are under "open" licenses (such as CC0 or PDDL), others are licensed for the beta itself, and every set is clearly licensed so you can see exactly what you can do with it.

Finding and exploring datasets is also core to the beta, so developers can search, browse and 'explore'. The searching and browsing are fairly familiar navigational terms, and don't need much explaining (though you can accurately refine your search), but Exploring is interesting. We've actually got a little "explore" button for each dataset, which gives you access to detailed developer documentation and the ability to browse the Linked Data structure of the dataset.

We've discussed Kasabi's API approach already, and the ability to access the Core Kasabi APIs, and develop your own are included in the beta. You're able to store your

SPARQL queries too, and share them with others, or hit them with a URL request and get data back. We've called this SPARQL Stored Procedures, and it's the basis of one of the APIs you can tweak and share.

We're currently working on our data publishing tools, which will be available during the beta period. Kasabi will let anyone publish datasets, and the raft of publishing and curation tools will be launched soon.

Kasabi Developer Network

When we first introduced the Kasabi Beta, it was to a room full of invited experts at a very special hackday. The very first non-Talisians to see Kasabi had taken the afternoon to fully get their heads around it, and hack away at it. In my hackday introduction, I asked for hacks and apps, visualisations, APIs, "surprises," and feedback.

I certainly wasn't disappointed.

In the span of a few hours, we had several working apps, visualisations of data, shared APIs and plenty of surprises and feedback. I'm running out of space here to do justice to some of the apps turned out, so I will point you at http://bit.ly/kasabi-hackday-hacks for more about a few of the tools that came out of the hackday.

More important than the interesting apps and demonstrations, though, was my first taste of Kasabi developer feedback. As the new Community Manager, I'm looking after the developers and the beta-side of the project, and this was my first chance to hear back from developers actually using Kasabi!

The feedback from the hackday was just the start of the help our beta testers have been in showing us how they find using Kasabi. For several weeks, we've released updates and improvements to the Kasabi beta, and each release has been influenced by the suggestions, bugs reported, and features requested from the testers. It has been an important time for the Kasabi project, and by the time you read this, even more will have been improved.

It has also been nerve-wracking to show off something we've been working on behind the scenes for the first time, but also exciting. The next phase of Kasabi will be a more public launch of the beta with a wider range of people able to look inside: to publish and explore datasets and customise and build on top of interesting information.

So, whether Kasabi is in closed or open beta by the time you read this, I would love to have more insight from Nodalities' readers. Kasabi has been built initially from the ideas of the team, and improved by the developer network. But it is very much part of our vision to include a wide range of people into the ecosystem. If you're interested in reading a bit more, you can find our blog at blog.kasabi.com. You can sign up for your own testing account at kasabi.com, and drop into the #kasabi channel on IRC (freenode.net) or follow @TeamKasabi on twitter.

Nodalities Magazine is a Talis Publication

ISSN 1757-2592

Talis Group Limited
43 Temple Row ,Birmingham, B2 5LS
United Kingdom

Telephone (within the UK): 0870 400 5000
Telephone (outside the UK): +44 121 717 3500
www.talis.com/platform
nodalities-magazine@talis.com

The 10th International Semantic Web Conference

ISWC 2011

October 23-27, 2011
Bonn, Germany

ISWC 2011

ISWC is the major international forum where the latest research results and technical innovations on all aspects of the Semantic Web are presented.

In Bonn, scientists from all over the world will meet to present novel and innovative research that demonstrates new trends and addresses scientific challenges for the Semantic Web.

Bonn is easy to get to being within a little more than one hour's reach from two international airports (Frankfurt am Main and Cologne).

Key Topics

- Management of Semantic Web Data
- Natural Language Processing
- Ontologies and Semantics
- Semantic Web Engineering
- Social Semantic Web
- User Interfaces to the Semantic Web
- Applications of the Semantic Web

Keynote speakers

- Sandy Pentland
  Massachusetts Institute of Technology (MIT), US
- Gerhard Weikum
  Max-Planck-Institute for Informatics (MPI), DE
- Frank van Harmelen
  Vrije Universiteit Amsterdam, NL

Main events

- Research track presents original, principled research papers addressing both the theory and practice of semantic Web technologies. Abstracts due June 16, full papers due June 23, 2011.
- Semantic Web In Use track presents papers describing innovative use of Semantic Web technologies in real-life applications. Papers due June 23, 2011.
- Doctoral Consortium provides an opportunity for doctoral students to present their current progress and future plans and to get advice on their work.
  Papers due July 11, 2011.
- Posters and Demos present the latest developments in the field and showcase Semantic Web systems. Submissions due August 15, 2011.
- Tutorials offer hands-on guidance on using key Semantic Web technologies. Tutorial proposals due May 6, 2011.
- Interactive Workshops discuss hot topics and new research areas. Workshop proposals due April 15, 2011.
- Semantic Web Challenge provides a forum for the best Semantic Web applications to compete.
- 2nd Linked Data-a-thon. The main goal of the competition is to develop innovative applications that showcase the benefits of Linked Data ... in two weeks!
- Industrial Track

Join the ISWC 2011 community

twitter
www.twitter.com/iswc2011

facebook
http://www.facebook.com/iswc2011

Platinum Sponsor

fluid Operations

Gold Sponsors

ontoprise
know how to use Know-how

poolparty

http://iswc2011.semanticweb.org

E-mail: confsec@uni-koblenz.de