# Nodalities

THE MAGAZINE OF THE SEMANTIC WEB

www.talis.com/nodalities

Issue 13

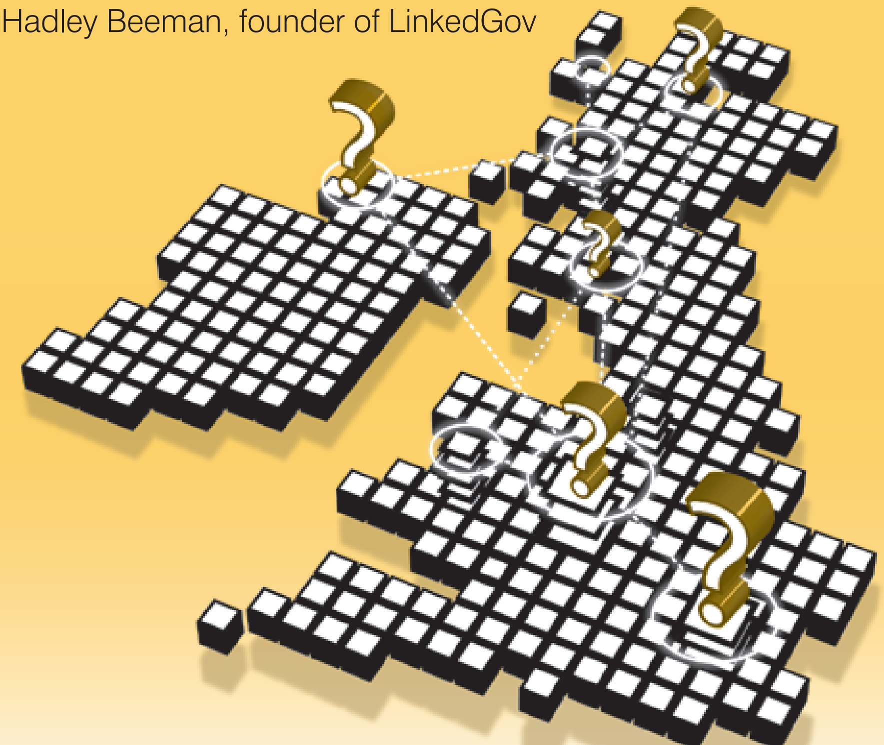

INSIDE



# The challenge with new tools:

## how LinkedGov is tidying the UK government's data

By Hadley Beeman, founder of LinkedGov



When Sir Tim Berners-Lee famously told then Prime Minister Gordon Brown that we should put all UK public data online, I, for one didn't appreciate what a mess we were creating for ourselves. Like a delivery of tools for a workshop that doesn't yet have workbench space for them, we have been happily flooded with data sets in a mish-mash of formats, structures and degrees of polish.

The creation of data.gov.uk has given us a hardware store's worth of tools, currently more than 5,000 data sets. As a data community, we are thrilled to see new data appearing by the day, not to mention new apps, visualisations and viable businesses built from individual data sets. When you've had neither tools nor workshops, getting tools is a good thing! Inevitably though, we find ourselves wanting to do more.


*continued on page 3*

**SUBSCRIBE NOW**

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

**FOLLOW US**

Follow Nodalities Magazine on Twitter: @nodalities

# Editor's Notes

Data changes the way things happen. Metrics inform decisions for investment, facts and assertions fuel industrial research, and mathematical and logical models allow for the expansion of human knowledge. Linked Data is playing an increasingly important part in the organisation and publishing of information. But the world of data is bigger than any concerns solely about data formats, frameworks, or the design of information architecture.

This issue looks at the world of big data through the lens of the Semantic Web. What we're seeing areas in need of development, as well as reports of important advances in the state of the art in Linked Data development. Nodalities hasn't shied away from discussing the challenges facing Linked Data in the past, and this issue surfaces several important calls to action for the community.

So, Olaf Hartig tackles one of the Linked Data world's biggest challenges—and most discussed topics—in his article on provenance. Laurian Gridinoc brings up a possibly neglected technology by discussing his vision for a more powerful browser.

Hadley Beeman's covering story discusses some of the challenges facing the Open Data movement. She also introduces her LinkedGov initiative that seeks to simplify the data curation tasks that are beginning to take root in public sector data projects.

The world of data and information is closely related to research and science, and I'm excited to be able to include an article from Kaitlin Thaney about the new Digital Science branch of Macmillan publishers. We've also got a story of discovery from Toni Bradford, who discusses her computing company learning the ins and outs of semantics through big ideas.

The Semantic Web community has proven itself more than capable of meeting challenges, and I'm genuinely excited to see what can be built when obstacles and difficulties sharpen development and thinking. As always, if you've got a story about Linked Data (or anything you find fascinating!) get in touch. You can get a hold of me (zb@talis.com) or follow @nodalities on twitter.

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

Continued from front page.

**By Hadley Beeman, founder of LinkedGov**

In the same way that new tools come with user manuals and DIY guidebooks, datasets are more useful with metadata and contextual links to other data. A significant amount of work is required to clean and reformat the data, link it (both internally and with other datasets on the Web), and add metadata to make it understandable to users outside the team or public body that published it. Many entrepreneurs are dissuaded from creating new businesses on this data, put off by the months of overhead necessary to prepare the data before they can even begin coding an application.

LinkedGov is a community project to make the data more useable and useful. We are breaking down a sizeable list of tasks into quick activities or games, distributing them across many people, and sharing the clean, linked public data that we produce.

## What is the problem?

Though it is wonderful to have data flowing from all parts of the public sector, we are running into a number of barriers to fully exploiting its potential. The distribution of people, the sheer amount of work required, the budgets available and the youthful maturity of the open data movement are all contributing factors. The Government's vision of new businesses, greater transparency and efficiencies in the public sector can't be realised until we can resolve some of these challenges.

Most of the data published from the various governmental and public bodies has been used in their day-to-day operations. Like borrowing a sander from a friend who has it set up for their own projects, the data sets often keep the quirks from the team who published it (as an example, we find many abbreviations and codes that are specific to them). These teams are impressively keen to help us – they use public data even more than most developers, and they would love our help to make it more useful – but as the public sector is facing cuts and efficiency drives, they have less resources and less time available to help. LinkedGov needs to make it quick and easy for them to help.

The next contributing factor is the challenge of standards. The drill bits for my power drill may not fit in your drill. Our parallel with data comes in publishing formats (RDF, XML, CSV, JSON, and .xls) and beyond that, we haven't agreed methods of organising data. For example, if I'm publishing data on electricity usage, should the rows be the addresses of buildings and the columns hours in the day? Or should the columns be kWh for an entire day? And once I've decided what I want, is that going to be the most useful standard for everyone?

Open data, as a discipline, hasn't yet reached the level of maturity where we can agree, for all topics, that full data standards exist and should be implemented. This means that LinkedGov must make use of what is the best (or most used) standard, but we can't just choose standards to impose on the community. We must also foster the creation of standards and the change process in implementing them, as the entire data community carries on working to nail down some of these important details.

The volume of data we are seeing (and anticipating) provides us with an enormous opportunity in linking it. If we have a data set of roads and potholes from a local authority, we can't search that data by postcode until we have linked the postcode database with the roads in the data. Fortunately, we already have great tools for linking data, where we can follow existing schema or ontologies.

*Most of the data published from the various governmental and public bodies has been used in their day-to-day operations. Like borrowing a sander from a friend who has it set up for their own projects, the data sets often keep the quirks from the team who published it.*

They will happily find connections between data sets and data types in those sets, but often need to check with a person: "Is this data referring to 'Pothole –the hole in the street' or 'Pothole - the band?'" As a one-off task, it isn't much, but when we extrapolate to the hundreds of thousands of datasets we are expecting from the various parts of the UK public sector - the workload is daunting for any one team or even company.

As Mary Poppins said: "In every job that must be done, there is an element of fun. You find the fun, and - Snap! The job's a game." The more fun we can make these tasks, and the more we can distribute them across lots of people, the easier the entire workload will be.

## What is LinkedGov?

LinkedGov is a project where we can tackle these challenges as a community. We are aiming to provide the tools that developers and researchers need: access to the UK's public data in a format of their choice, mapped to a data standard that works for them, with enough metadata to be understandable outside of the public body that released it.

To make this happen, we are working to break down all the tasks into 30-60 second activities that require a minimum of effort from any user to provide a huge amount of value. We will take government data sets in any data format, mark problems or questions in the data, and queue up those questions for civil servants, local government officers and NHS managers who can help. We will break down the linking tasks into those 30-second bits and arrange them into games and activities so that a crowd, even those without government experience or data familiarity, can help us make this data useable.

For LinkedGov to be useful to as many people as possible, we are seeking input and expertise from many areas in our community. We need to be a confluence of expertise in many areas (linked data, XSLT, data formats, games, data standards, Transparency policy for the UK government, the machinery of government and the public sector). To be an attractive proposition for every one of those groups, we are building tools and rewards to offer each of them for their participation: for the developers and data experts, we are offering clean data and open source code. For the non-techies, we are creating a Question Site, where a user can interrogate the data for a simple answer to a question without needing to write an SQL or SPARQL query. We aim to build in enough feedback loops to ensure that LinkedGov gets the mass participation it needs and delivers what the community needs – staying on top of those needs as they may change.

By engaging all these groups and creating tools to meet their needs, we are breaking down the load to manageable pieces to release the crowd. Rather than become overwhelmed with the amount of work involved, we are tackling them as a community, making it easy for everyone to contribute.

This should provide us with a full workshop of tools, available to anyone, with clear information on how to use them. With our tools at the ready, we can focus on our projects: building apps, research projects and new businesses from government data.

*LinkedGov is sponsored by IC tomorrow, a programme run by the UK government-funded Technology Strategy Board. We are actively encouraging new input from the community – please join us at LinkedGov.org.*

# Digital Science

**By Kaitlin Thaney, Manager of External Partnerships for Digital Science, a new technology company out of Macmillan Publishers. She comes from the open science world, most recently managing the science division of Creative Commons in the US.**

**This past December, Macmillan Publishers launched 'Digital Science' - a new company dedicated to bridging the gaps between the analogue and digital research worlds through tools, services, and better use of technology.**

This article will take a closer look at the ecosystem in which Digital Science exists, and delve into our initial areas of focus, areas of the research process we know can be enhanced through smart, scalable tools. Our goal: to make research more efficient through the better use of technology, leading to faster discovery, more comprehensive research management and, in the end, better science.

Science and technology have long been inextricably linked. The first scientific "computing machine" - the Antikythera Mechanism – is a remarkable example of this, an ancient Greek device that through a series of computations could predict lunar eclipses. This is an early example of how technology enhanced and enabled discovery in a field -in this case, astronomy. Similar inventions have dramatically changed the way disciplines tackles certain problems, whether through advances in instrumentation (Hubble space telescope, DNA sequencing devices, particle accelerators) or even data storage and distributed computing.

The introduction of the Web continued this era of advancement, transforming the way humans and machines process information. The Web has fundamentally altered the way we think of and interact with entire sectors, from commerce, to education and the arts.

Science is no exception. The web brings a wealth of opportunities to innovate on the scientific process and change the way research is done. We have the tools and technology now to accelerate discovery in a way never seen before and make research far more efficient. But we're not there yet. Gaps remain, and Digital Science hopes to bridge them.

*"The future is here ... just not evenly distributed yet."*

- William Gibson

William Gibson's quote resonates particularly well when it comes to thinking about the opportunities presented to science by modern technology. For more than 20 years, we have seen deeply-rooted processes altered by the existence of the World Wide Web. The way we buy and sell, teach and learn, search and communicate. But, for all of the benefits and uptake seen in e-commerce, fine-tuned search or the way we manage our music collections online, those efficiencies have yet to gain full traction in science.

*Humans are no longer the sole consumers of online information. Scientists now rely on machines to help carry the weight of heavy experimentation and discovery – even down to the basic function of how they search.*

The move to digital has completely changed the way we generate, publish and distribute scientific "knowledge", forcing us to think of new ways to innovate online. It has changed the way scientists operate in laboratories, from knowledge exchange and collaboration, to acquisition of materials and running of experiments. It has even (and in many cases, most obviously) transformed how they sift and navigate through the treasure trove of online content, and more and more datasets.

And humans are no longer the sole consumers of online information. Scientists now rely on machines to help carry the weight of heavy experimentation and discovery – even down to the basic function of how they search.

Better knowledge discovery and tools are needed if we are to help scientists render real value in an increasingly connected research domain. Our aim at Digital Science is to support that and address some of the main blocking points with state-of-the-art software and tools.

We've identified a series of target areas, each with unmet needs and technology challenges where we are certain we can help.

Our focus areas can be broken down into three categories: knowledge discovery, software applications, and tools for better research management. We're researcher-focused in our approach, but also understand that research administrators and funders are a fundamental player in the scientific process, and they also could benefit from smarter tools for decision-making.

We combine in-house expertise and development with external partnerships and acquisition to help build our product portfolio, working with an array of start-ups, established companies, individuals and other collaborators. This approach is deeply rooted in the belief that reinventing the wheel is not the most productive use of time, and that partnering with other tool developers to bring their work under the fold and help them develop their ideas is often much more useful.

I'll explain more about that area later on in this piece. As mentioned previously, we also are aware that we're not just building these tools with humans in mind, but also machines - and ensure that our tools are interoperable and built to scale.

### Knowledge Discovery

We're making steps towards better annotation and search. But we are still quite a way from asking difficult questions of large amounts of information and rendering comprehensive, yet sufficiently sifted, results. Efforts have continued to spring up in recent years, but the technology remains clunky, expensive, and a strain for most to understand.

Our text mining business is built on the company SureChem, a chemical text mining company we bought in late December 2009. Their service searches patent literature and content in Medline for chemical structures, a space that to date has been plagued with a name disambiguation problem, making it challenging for scientists to quickly search for the chemical compound upon which they are basing their research. SureChem's technology tackles that issue in a novel way, serving as a search interface for chemical compounds where the user can query through both drawn structures and chemical name search. This is especially important for fields such as cheminformatics and drug discovery, where name disambiguation is an everyday challenge, and chemical structures are the best means to fine-tune a query.

SureChem represents the starting point on which we hope to base our text mining and knowledge discovery efforts, with plans to extend beyond chemical compounds. And to ensure that along the way, we're building these tools in a fashion that echoes SureChem's approach - streamlining the technology to be harmonious with the way in which the scientist think and their everyday practice.

## Software applications

We've talked thus far about "scientific information" largely in the context of scholarly content and patent literature, which we can now search, disseminate, annotate, and publish in new ways. But, as any lab researcher can attest, those are by no means the only materials or building blocks involved in the scientific process. From datasets to tracking and ordering information for biological materials (think cell lines, plasmids, lab mice) and devices, there are more parts to consider when it comes to better managing scientific information in the hope of making research more effective.

## Let's take the laboratory, for example.

The "information" in this respect often lives in the mind (and on the desk) of the lab manager, tucked away in myriad Excel spreadsheets, making it difficult to extract, let alone efficiently keep tabs on. There are expiration dates (just like jugs of milk, cell lines don't last forever), location information (which freezer is X in?), parameters for experiments and calibration information for equipment. Add to that the information attained from the experiment itself, the surrounding research and analysis often shared between collaborators internally and externally, and the problem is only exacerbated.

BioData, one of the start-ups we are working with, is tackling this problem. Their web-based tool provides an easy and affordable means of tracking projects, protocols, biological materials and inventory all in one place, creating a one-stop-shop for all of the crucial information needed in a modern lab. The tool was crafted by a scientist, frustrated with inefficiencies of his own work environment and unimpressed by other more costly lab information management systems out there.

We also plan to provide other tools and services to help ease day-to-day information management and increase researcher productivity.

## Research Management

From the beginnings of science, scientists have been evaluated in some way or another, be it through qualitative or later, more quantitative measures. Whether it is performance-based, citation-based, or any other host of indicators, scientists are now measured and rewarded based on such metrics. Each of these research measurements plays into the broader social issue of rewards and incentives, in many ways a key driving factor (like it or not) of scientific research.

The social issue of science — in terms of rewards and incentives, promotion and tenure, or even merely tool adoption, is a thorny matter to navigate. This side of the research process — research administration and management — is riddled with inefficiencies, and deserves acknowledgement as a constituency that could also benefit from better technology. Just like scientists at the bench, university administrators are hampered by lack of state of the art tools that can not only capture all the information streams that now make up the digital research workflow, but compile them in a fashion that's as comprehensible as it is comprehensive. From the point of view of university administrators and funding agencies, it's not only about providing individual credit but also about establishing rewards and incentives for activities such as collaboration that serve the interest of science as a whole.

> *"We're experiencing a Cambrian explosion of metrics"*
>
> Johan Bollen

There has been a drastic spike in the number of measurement systems for scientific research in the past few years, making this space even more complex. From traditional impact factor and citation based systems to more recent additions such as the h-index, Eigenfactor, or even "betweenness centrality", the myriad of systems are still each heuristic and imperfect in their own ways, each built to reward slightly different behaviour.

We can only make better decisions with better data, and we're working towards that through in-house development and also through a partnership with a London-based company, Symplectic. Their flagship product 'Elements' assists research administrators and academics in accurately capturing the research output for their institution, a task that's often heuristic in nature, imperfect in practice and difficult to implement. The system gathers a wide array of information necessary for research assessments and reporting, from research publications from faculty to funding streams and organizational structure.

The existing series of systems is imperfect, at best, a challenge to harmonise, track, reward, and even still measure. Each of these systems is based on citation analysis of the peer-reviewed literature - a useful measurement to have - but by no means the most comprehensive. Add to that the fact that these metrics are the basis for promotion and tenure committees, and the problem is only exacerbated.

Our aim is to provide research administrators with a more holistic view that helps to track various components of the digital workflow, such as sharing of data and resources. But first, we have to put the mechanisms in place to map what is already out there more effectively, and Symplectic is helping us do that.

## Next steps

We're still much nearer the beginning of this journey than the end. For all the recent technological advances, we're still a way away from fully realising the potential of information technology to accelerate the discovery and application of scientific insights. The research process is facing a sea change: a new era of digital technologies that enhance, expand and challenge entrenched scientific norms and practices. Digital Science was crafted to help provide tools to spur and aid this transformation, by creating software that understands science … and scientists.

*For more information about Digital Science or to inquire about how you can work with us, contact me at k.thaney@digital-science.com.*

# Integrating Provenance into the Web of Data

**By Olaf Hartig, research assistant with the database research group at Humboldt- Universität zu Berlin. He is working on novel approaches to execute queries over the Web of Data, including an assessment of the trustworthiness of query results.**

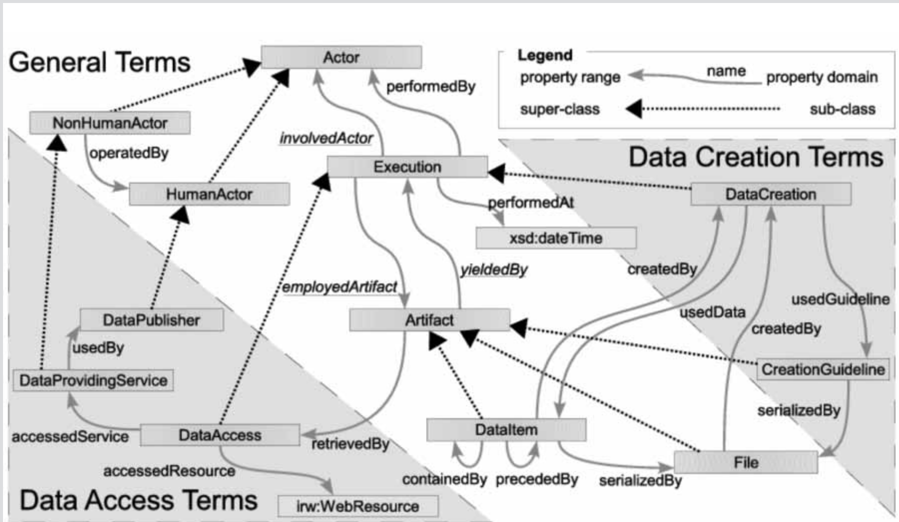


Fig.1

Today, a large amount of RDF data is published on the Web; large datasets are interlinked; new applications emerge that utilize this data in novel and innovative ways. However, the openness of the Web and the ease to combine Linked Data from different sources creates new challenges. Unreliable data could dominate the result of queries, taint inferred data, affect local knowledge bases, or may have other kinds of negative or misleading impact on applications that consume the data. Hence, questions of quality and trustworthiness must be addressed.

Assessing a data object with regard to its quality and trustworthiness includes considering the following questions:

- How was the creation of the data conducted?
- Who and what participated in the creation?
- Based on what input was the data object created?
- When was the data object (and possible source data) created?
- What happened to the data object since its creation; how likely is a manipulation?

These questions refer to the provenance of the data object.

We understand the provenance of a data object as everything that is related to how the object in its current state came to be. Hence, provenance information about a data object is information about the whole history of this object. This history may start long before the object has been created itself because the provenance of source data used for the creation is also a relevant part of this history.

In addition to quality assessment of data objects, provenance information has multiple other uses. For instance, provenance is crucial to understand the legal rights to use a dataset that has been generated based on data from multiple other datasets. Provenance information may also be used to reproduce a data object or to re-execute a certain process such as the collection and integration of data from a variety of sources. For an overview on use cases and a state-of-the-art understanding of different dimensions of provenance, we refer to the final report[1] of the Provenance Incubator Group who recently studied the area of provenance for the W3C.

To make use of provenance information in an application, this information must be obtained in the first place. We distinguish two main sources for such information: provenance recorded by the application itself and provenance-related metadata published by the providers of data. In applications that consume Linked Data from the Web, the first of these two options can only provide for a limited amount of provenance information. Hence, to obtain more complete knowledge these applications rely on provenance metadata from third parties such as the data providers.

To allow for a successful consumption of provenance metadata we conceive it as an absolute necessity that this metadata becomes an integral part of the Web of Data. This is only possible if the publication of this metadata adheres to the same principles that are used for the data itself. Therefore, we develop a vocabulary that allows providers of Linked Data to describe the provenance of their data using RDF. Ideally, these descriptions have to be created and made available as Linked Data, again. To initiate such a practice we extended several Linked Data publishing tools with corresponding metadata components.

### The Provenance Vocabulary

The development of the Provenance Vocabulary[2] is motivated by the need to describe the main aspects of provenance of data consumed from the Web. We distinguish two main dimensions of provenance that are typical in this context: data creation and data access. Some, more general concepts, such as actors, processes and data items, are relevant in both these dimensions. Consequently, the Provenance Vocabulary consists of three parts: general terms, terms for data creation, and terms for data access.[fig.1]

These terms allow users to describe all common cases related to Linked Data publishing such as: manual data creation, Linked Data interfaces over native RDF stores, wrappers over relational databases or over Web APIs, file-based data creations, data changes, retrieval and provenance of source data, publication responsibilities, and service operators.

[Fig 2] illustrates a simple example of such a provenance description: it describes an RDF

```
[] rdf:type prv:DataItem , rdfg:Graph ;
   prv:createdBy [
          rdf:type prv:DataCreation ;
          prv:performedAt "2011-02-06T16:38:05Z"^xsd:dateTime ;
          prv:performedBy <http://example.org/OlafsLinkedDataServer> ;
          prv:usedData [
                  rdf:type prv:DataItem ;
                  prv:retrievedBy [
                          rdf:type prv:DataAccess ;
                          prv:performedAt "2011-02-06T16:38:05Z"^xsd:dateTime ;
                          prv:performedBy <http://example.org/OlafsLinkedDataServer> ;
                          prv:accessedResource <http://en.wikipedia.org/wiki/Berlin> ]
                     ]
              ] .

<http://example.org/OlafsLinkedDataServer>
      rdf:type prv:NonHumanActor ;
      prv:operatedBy <http://olafhartig.de/foaf.rdf#olaf> .
```

*Fig.2*

graph which was created on Feb. 6, 2011; the creation was performed by a system that was operated by me; moreover, the creation included the use of source data, which that the system retrieved during the creation process by accessing the Wikipedia page about Berlin.

We develop the Provenance Vocabulary with comprehension and usability in mind. For this reason we apply a consistent scheme for property names and we provide properties that present shortcuts for very common cases such as file-based data creations. Another good practice for Linked Data vocabularies is the interlinking of related terms between vocabulary definitions. Such “schema-level links” improve the degree to which published data is self-describing. The Provenance Vocabulary adheres to this practice. For instance, the Actor class is defined to be equivalent to the Agent class in the well-known FOAF vocabulary. This relationship enables a FOAF-aware application to deal with actors mentioned in a provenance description as if they were FOAF agents, e.g., in visualizations.

To allow for a wide range of applications the vocabulary does not prescribe a specific level of granularity by which provenance has to be described. For instance, a data item in a provenance description can be a whole dataset, a specific RDF graph as provided by a Linked Data server, or even a single RDF triple.

While our vocabulary provides a basic framework to describe the provenance of data from the Web, it does not aim to support the description of every aspect and detail of provenance. In particular, to provide a more detailed account of certain data creations it is better to use more specialized vocabularies. We provide a comprehensive documentation in which we also discuss how other vocabularies can be used together with the Provenance Vocabulary.

*To make use of provenance information in an application, this information must be obtained in the first place. We distinguish two main sources for such information: provenance recorded by the application itself and provenance-related metadata published by the providers of data.*

### Metadata Extensions for Linked Data Publishing Tools

The integration of provenance-related metadata into the Web of Data can only be successful if generating and publishing this metadata is possible with minimal effort for data providers. For this reason, we developed metadata components for widely used Linked Data publishing tools, including Triplify, Pubby, and D2R Server. These extensions provide a mechanism for adding metadata to the RDF graphs served by the publishing tools. Hence, with our metadata components the tools publish metadata together with the

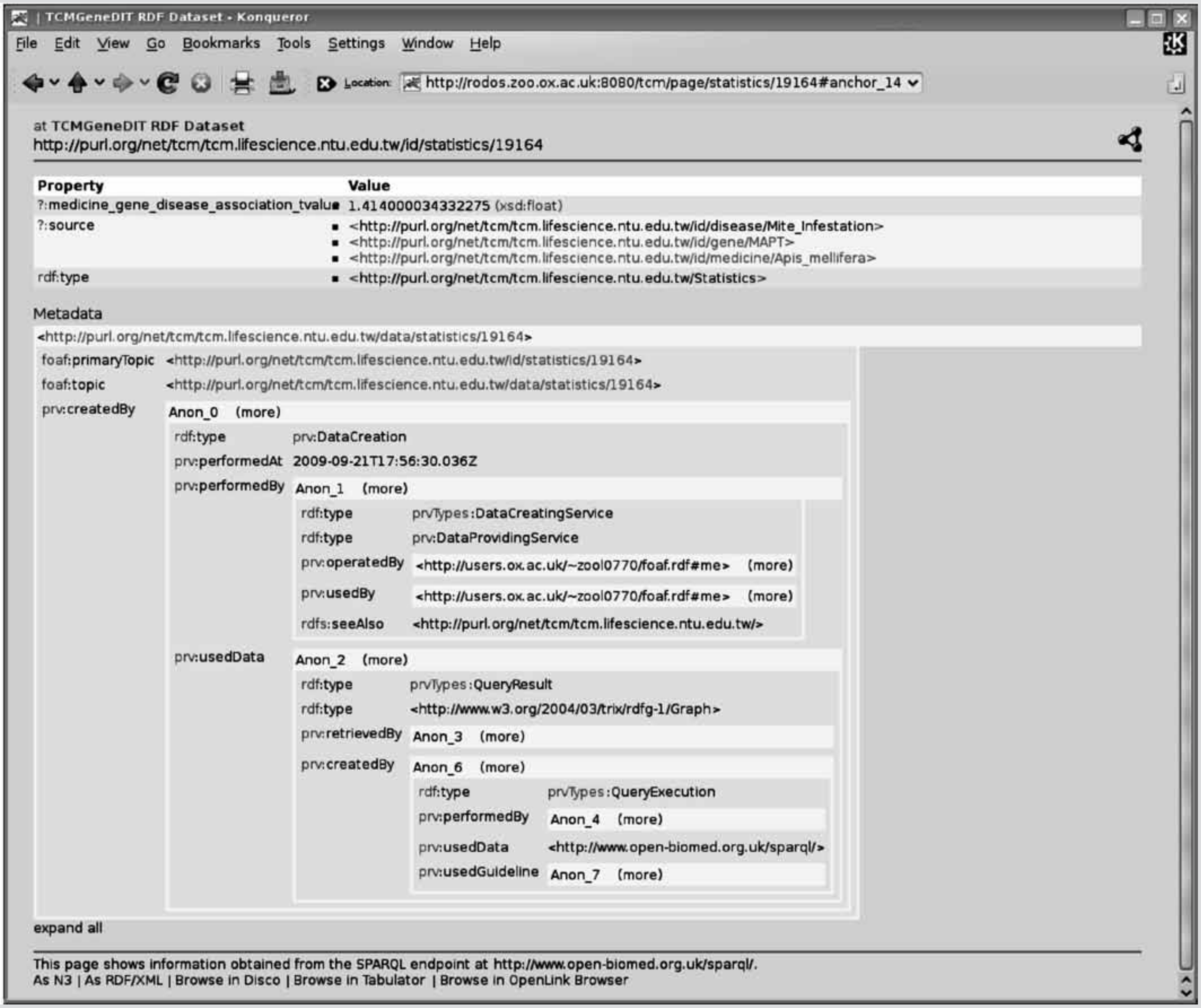



*Fig.3*

data itself. Additionally, the provided metadata is also visible in the human-readable output of the publishing tools.

While the developed components can be used for any kind of metadata, we added a default template that -out of the box- provides various pieces of provenance-related metadata, represented with vocabulary. Data providers can easily enrich the metadata generated by default: they just have to configure a few parameters, such as the name and URI identifying the provider or the URI of the dataset. Nonetheless, it is also possible to adjust the metadata template itself. Hence, our components provide a flexible metadata generation framework that can be customized for every data source with minimal effort. With data providers upgrading their Linked Data servers to the latest release of these tools the amount of provenance information added to the Web of Data will increase significantly.

> *To make Linked Data more reliable it is crucial that the publication of verifiable provenance information becomes a best practice for data providers.*

While the Provenance Vocabulary and our metadata components are an important contribution to the integration of provenance into the Web of Data, we understand them as a first step. To make Linked Data more reliable it is crucial that the publication of verifiable provenance information becomes a best practice for data providers.

**References:**

[1] http://www.w3.org/2005/Incubator/prov/XGR-prov/
[2] http://purl.org/net/provenance/

# Linked Data and Scholarship in the Humanities

**By Sarah Bartlett, Senior Analyst, Talis**

In a context of emerging interest in Linked Data across academia, it is useful to start to examine its potential in areas of scholarship not traditionally associated with technology. To date, work on the Semantic Web seems to have been centred on the physical and social sciences. But literary textual relationships open up a whole new spectrum of relationships, and offer insights into the power of Linked Data.

It was Julia Kristeva, a French literary theorist, who coined the term intertextuality in 1966. For Kristeva, the literary system acts like a collective mind that writers unconsciously draw on. Writers' minds, then, contain every text they've ever read, any utterance of which might surface subliminally or explicitly in their writing. This presents obvious problems in terms of identifying and mapping intertextual relationships; such approach would result in an unverifiable model of almost infinite size.

Into this unwieldy situation steps Gerard Genette, who urges literary scholars to focus on demonstrable and identifiable relationships. Specifically, he establishes a typology of relationships to describe different relationships between literary texts. This could form the basis of an ontology:

**Architextuality** – This links the text to a grouping such as a literary genre.

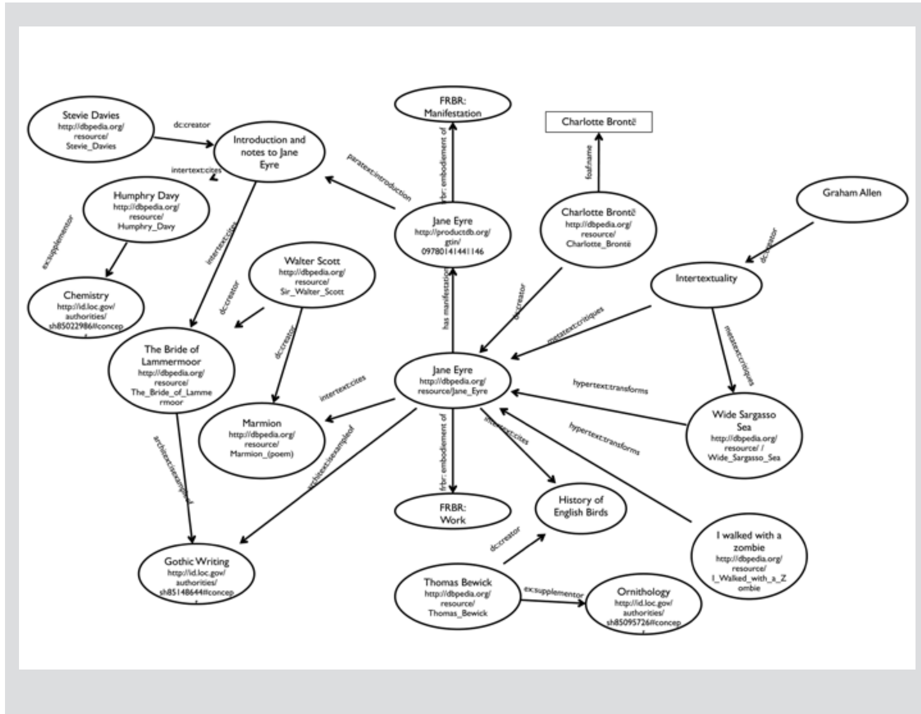
**Intertextuality** – The presence of a text in another text, typically in the form of a quotation, but also plagiarism, or allusion, with the exception of commentary.

**Paratextuality** - Elements which lie on the threshold of the text, such as titles and chapter titles, introductions, notes and book covers.

**Metatextuality** – When one texts presents criticism of or commentary on another text.

**Hypertextuality** – The transformation of an entire text into another, such as Homer's Odyssey to James Joyce's Ulysses. Translations, prequels, sequels and parodies all fall into this category, which is as broad as it is fascinating. It is useful, then, that Genette provides further layers of categorisation that may be used ontologically.

In the RDF reproduced here, I apply Genette's framework to a subset of the relationships around Jane Eyre by Charlotte Brontë, a novel that happens to be rich in intertextual references. My aim is to demonstrate to people unfamiliar with Linked Data principles how semantic technologies could transform scholarly understanding of intertextual relationships.




Besides the obvious application of modelling critical studies of literary works, for example critique of Jane Eyre in Intertextuality by James Allen, other areas of this model also merit scholarly attention.

Firstly, by linking Jane Eyre to the Library of Congress subject heading of Gothic Writing, we can start to group together all the instances of a genre, identifying intertextual influences within the grouping. This data could then be reused in interesting ways – a timeline would represent the development of the genre visually, for example.

A similar approach can also lead to important insights at an interdisciplinary level. In the top left corner, we see that in the introduction to the Penguin Classics edition of Jane Eyre, Stevie Davies contends that Humphry Davy, the 19th century chemist, was a significant influence on Victorian authors such as Charlotte Brontë. Brontë drew heavily on his ideas to represent physical attraction and repulsion between her fictional character. By linking up to the Library of Congress subject heading Chemistry, we can start to trace the influence of non-literary disciplines in fiction, as scientific fields enter the popular consciousness over the course of that century.

A network model pointing to authoritative URIs also offers a robust resource for scholarly verification. In this instance, the intertextual relationship between Jane Eyre and Wide Sargasso Sea (a prequel to Jane Eyre written by Jean Rhys) is validated by the co-presence of both novels in the critical study Intertextuality by Graham Allen.

At an authorial level, we can also see the influence of Walter Scott on Brontë with the model showing that Scott's poem Marmion is cited within the text of Jane Eyre itself. Furthermore, Davies's introduction to Jane Eyre cites Scott's Bride of Lammermoor as another example of the gothic novel.

Consistent with other exemplars of the Linked Data approach, this model, although only a partial picture of intertextual relationships around Jane Eyre, is already delivering important insights, which will increase exponentially as further data and relationships are added. The graph is always incomplete, but always extensible and useful from an early stage. And in an area of scholarly endeavour such as intertextuality that is not always clearly and objectively defined, the traceability that Linked Data brings makes the provenance of ideas and theory more discoverable.

*Sarah Bartlett gratefully acknowledges the support and guidance of Dr Bill Hughes, University of Sheffield.*

# The Exam Question

**Toni Bradford, founder of AB Computing Limited, specialising in Data Modelling and Business Analysis. She is now heavily involved in the Semantic world which has brought her back to her original love of natural language processing.**

No one can ignore the impact that the Semantic Web is having in the world at large. Even my elderly relatives, who have never touched a computer, ask frequently 'Can you find me XYZ on the net?' They don't know that it is semantic or a web. However they perceive it, they feel its power.The semantic web is pervasive in a variety of areas, but so far it has not made a serious dent on the financial and commercial powerhouses of the world and we need to ask why. My view is that the technology requires a couple of tweaks to make it truly usable at a commercial level and that we lack imagination.

Our background is in standard information systems. Systems of all types, sizes and shapes, but with a constant: we have always used mainstream technologies. We keep abreast of what is going on, but for serious implementations we would only use what is generally considered 'safe'. A casual conversation changed all that.

The starting point of our journey into the Semantic world was almost three years ago with a chance comment from the representative of a foreign country chatting about the possibility of building a security system for his homeland. As consultants we have been working for the last 20 years on many UK government systems so we viewed the potential request with the appropriate respect, but without undue fear.

The project did not progress beyond a proof of concept, but after a while we were hooked. We began to understand the capabilities that Semantic technologies could deliver in a commercial context and decided to invest on the development of tools and techniques which we believe have a serious commercial future.

Following on from our chat we started dreaming, thinking, planning and designing, and came out with 'a vision'. We did not have a formal statement of requirements to impose boundaries on our system dreams, and whenever we asked questions about what data we could use, the answer was always 'You can have whatever you need'. This opened up a set of possibilities that would be unthinkable in most European countries.

We started considering the implications of utilising data from Border Controls, Immigration, Penal system, Births-Deaths and Marriages, Finance, Banks, ATM, Phone Traffic, Supermarket car parks, Traffic Cameras, Hospital admissions, etc., etc., etc. The richness of the questions that could be asked exceeded our wildest expectations.

Our 'vision' encompassed the possibility of multiple nations using the same architecture, and making the capabilities available not only within one country, but potentially across the world.

At the heart of the system would be a huge database capable of storing new data and able to utilise legacy information. The vision included the creation of 'brotherly data links' between departments in one country and 'friendly data links' to other governments. We created a concept where each installation could define the level of trust with every other installation, and control the degree of data sharing at a very fine level of granularity. We described this as 'Soft' and 'Hard' Federation.

*Our 'vision' encompassed the possibility of multiple nations using the same architecture, and making the capabilities available not only within one country, but potentially across the world.*

The database would receive feeds from a wide range of data sources, and have the usual mechanisms to push the data in and orchestrate its use. Data would be qualified by its provenance and degree of reliability.

Two other components of the system were a heavy duty Events Engine and an Inference Engine. The Events Engine would understand the relevance of certain event instances and event sequences; and be able to send action requests when the temporal rules fired. The Inference Engine would be a consumer of the data qualifications and be driven by rules. It would scan the data continuously and be our 24x7x365 detective.

The external face of the system, the GUI, had to allow detectives to enter data in pictures and to describe, also in pictures, the questions they wanted to ask.

The GUI would also have to make it possible to create 'named' relationships between the data and to formulate queries about the data stored and about the relationships: Relationships had to be the 'objects' in their own right.

The system had to work on the desktop and on mobile devices; it had to be intuitive and accessible to IT novices, but provide answers to complex queries.

Once we created this list we started thinking about how we could make it happen.

We knew that the Relational model would not be rich enough to express the complexity of the requirements and would be incapable of defining the relationships as first class objects. We had to look for alternatives, but given the type of system, whatever we chose had to be of industrial quality.

Initially we looked at the Object model and the capabilities of Object Databases. While they offered sophisticated mechanisms for storing our type of data, they did not give us the necessary richness to formulate the queries. We had to keep searching.

Then we looked at RDF databases. We felt that they should be capable of working with seriously large volumes of data, but we needed to be sure that we could represent our universe so started modelling the ontology and making experiments.

We soon found that RDF databases provided the data flexibility and the querying capabilities we needed, but SPARQL worked at too low a level. We needed the usability of the relational model with the flexibility of the semantic database. This was a blow since we had been very hopeful. We decided to list our requirements and the RDF shortcomings, to determine whether we could bridge the gap.

We needed to group attributes and form a coherent 'semantic entity', which represented our understanding of that object within our universe. The RDF databases didn't have it without significant programmer effort. Querying the data in this system is of paramount importance, but working at the SPARQL level would be very cumbersome. We needed to simplify the means to query the data without using SPARQL directly. Another essential requirement was enforcing data integrity. Given the nature of the system it was essential to impose strict control on the data entered into the database. RDF databases are flexible, but very open. We needed to be more prescriptive.

We definitely wanted the inferencing capabilities that came with the semantic model.

We came to the conclusion that we could get what we needed by placing a layer between the GUI and the RDF database. This would give us the ability to define and manipulate higher order objects, and translate user queries into complex SPARQL.

We built a Semantic Entity Layer and developed a query language in XML that translates the requests from the user interface into the appropriate SPARQL query(ies). The resulting SPARQL is sent to the RDF engine where it is executed. Any results are returned to the Layer, where they are repackaged and sent back to the graphical application. In the Semantic Entity Layer we built a Search Engine with algorithms that allow us to formulate queries with flexible links, extending the concepts of path expressions that are still not fully standardised in SPARQL. By using the layer we are able to express within the GUI 'simple' queries which would be long and complicated in pure RDF([fig1]).

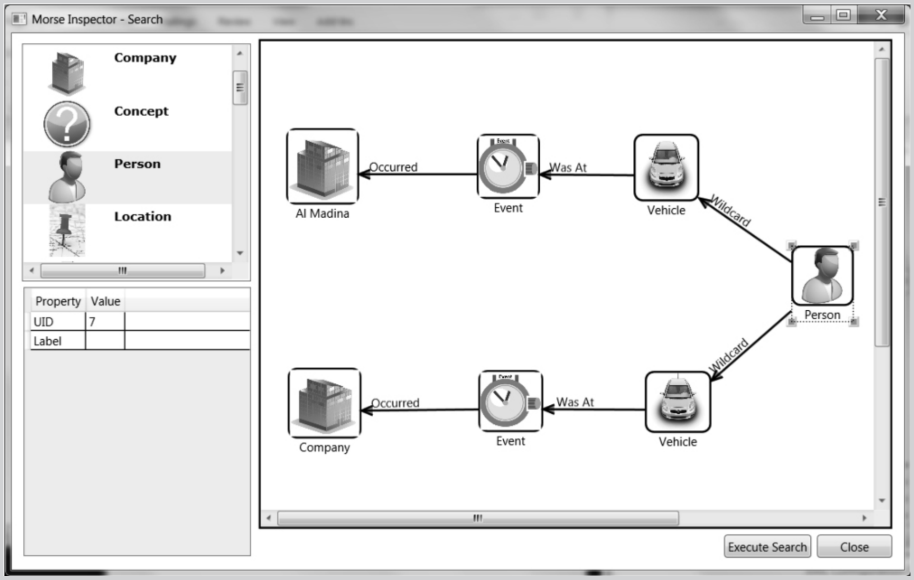



Fig.1

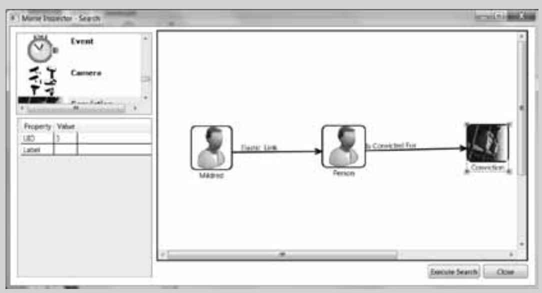


Fig.2

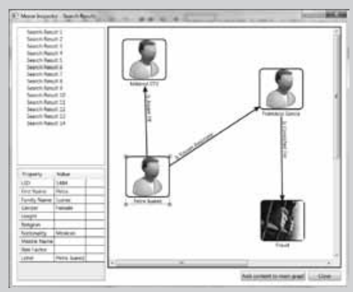


Fig.3

This query could be saying: return all the matches for VEHICLEs that were captured by the camera of a COMPANY of Type 'Petrol station' on <<Date>>, where the car was owned by a PERSON of <<fieldname_value>> and the VEHICLE was also captured by a camera at COMPANY 'Al Madina' on <<Date>>.

We are also able to phrase a query such as "Return all the PERSONs linked within 3 hops to a PERSON with a first name of 'Mildred' and who have been convicted for 'Fraud'"([fig2]).

One of the results returned, in our universe would be as seen in [fig3]

The Semantic Entity Layer allows us to create, manipulate and query 'Semantic Entity Types' to suit any business domain, PERSON, LOCATION, VEHICLE, INVESTMENT, POLICY, ACCOUNTS, etc., and to create, manipulate and query the relationships between the Entity types.

Our early models were very static, so we introduced context objects – INCIDENT and OBSERVATION to define a start time and duration, and to link multiple other entities.

Our product roadmap now includes plans to exploit these context objects through temporal animation within a GIS system to help the user visualise the development of the stored data in time and space.

We are convinced that the power of Semantic technologies will gradually take their place in the mercenary world of commerce, but the industry will have to create the tools to allow semantic data to be used at a higher level and give it greater usability.

We see Semantic applications being used in Investment for the transparency and traceability needed to avoid fiascos such as what happened with the Toxic Investments, where people had lost track of what was linked to what; proactively identifying gangs involved in fake vehicle crashes and Insurance fraud; helping to identify potential marketing strategies or providing the traceability of food products. The only limit is our imagination.

# A Semantic Browser

**Laurian Gridinoc is a Creative Software Developer at Talis Education**

The past years have brought wonderful developments for the web of documents: HTML5, CSS3, web fonts etc. The same is true for the web of data: the LOD cloud has enabled access to an immense amount of meaningfully interrelated data.

However, through one browser these two webs feel disconnected: one is sparkling, the other one looks like a ticklish graph with esoteric labels at best. And that's fine: one is for us; the other is for the machines.

Of the all machines that tread on those webs, one is usually forgotten: your very browser, a brilliant piece of software engineering that can handle a mix of old and new standards and deliver your shopping experience, your friends' activities, etc. Yet, it cannot directly handle linked data in a usable way.

Ever since the advent of the semantic web, researchers forecasted the coming of the semantic browser, a browser that will break free of the syntactic links (the ones "hardcoded" in a web document) and dynamically present semantic ones, enabling alternative uses of the viewed document.

One of the first semantic browser experiments was Magpie, created at Knowledge Media Institute (KMi), which would allow the use to select a specific domain to be applied to the viewed document. Upon selection, Magpie will load the domain ontology, recognise various instances of the domain concepts, and provide concept specific services.

One of the limitations of Magpie was that it delegates to the user the choice of the domain, of the use case; this was a rather technical limitation which disappeared today with the LOD cloud.

Nowadays the same augmentative functionality is achieved by various tools, running as third-party scripts embedded in the documents or even as browser extensions. These tools use Natural Language Processing (NLP) techniques to recognise entities within the document and anchor them (with a degree of ambiguity) into the LOD cloud, providing specific actions based on the type of the entity and its connections through the LOD cloud without requiring the user to choose a-priori an use case.

Such tools are quite helpful in making bare content actionable, but lately, bare content is seldom found online. Now content is framed by various options covering the most frequent use cases, and applying an additional augmentation tool floods the user with almost the same options as these tools are based on recognising entities within a page and not what that very page allows you to do with them further.

We used to bridge the gap between these two webs with NLP tools, but it was just a matter of time that the web of data will permeate through the one of documents, it happened first with microformats, then with RDFa. Sadly, one of the most important incentives for RDFa is search engine optimisation (SEO), but I believe that in time the most important incentive will be what the browser will enable as experience on RDFa annotated sites.

in that context creating in the end a poor user experience. A semantic browser could filter out and provide replacements to those actions if that page would be properly annotated.

Now, research done on Magpie showed that users don't like to select what they want. How can a semantic browser know better than a site what the user wants? The answer lies in the difference of perspective each has. A site may know little about the user but a lot about its products and services, generic buying behaviours, reviews and so on; while the browser can infer a multitude of things from the browsing history, a meaningful history of pages with annotations, from the currently open tabs and windows including your very own social network.

*There are so many things that can be connected through the LOD cloud, that you may say that there are too many possible use cases and no one will create a semantic browser that will cover all of them.*

Such annotated sites can be easily used in mash-ups that will compare and contrast every little detail of their offerings, allowing users to weight the prices and the quality of service. But this is more of a business to business scenario. There are things more important than competitive prices to the user, and that's the shopping experience.

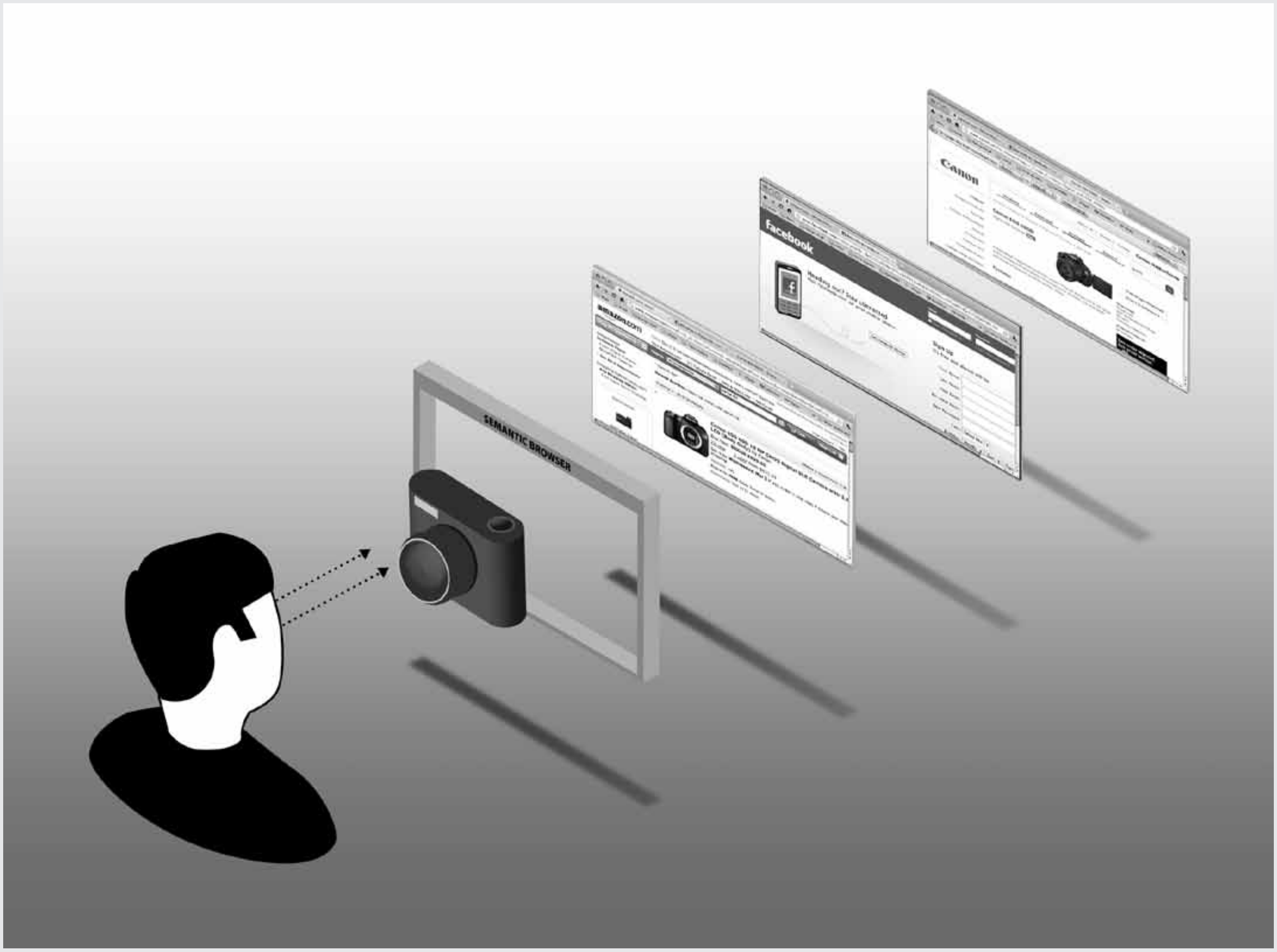
Consider that the aforementioned semantic browsing tools could now be less based on NLP and more on authoritative metadata. This means less ambiguity in enabling alternative use cases, and moreover, with vocabularies like GoodRelations we will be able not only to know what is what, but what we can do with it, therefore not just showing similar options but enabling only one and providing a justification behind that choice.

A web site offering a product or a service is designed to cover a set of specific use cases. When you visit a product page almost half of it is irrelevant to your intention as that page presents you with everything that is actionable

Let's consider this rather simplistic scenario: I have three tabs opened in my smart browser, one on Amazon looking at cameras, one on a review site for a specific camera and one on Facebook where I'm chatting with a friend.

The browser could pick up the name of another friend mentioned in the chat, discover that his birthday is next week and enable gift wrapping suggestions on Amazon. On the review site of a specific camera, next to sample images I will have pictures taken by my friends with the same model, discovered through Facebook. And since Amazon sells that camera, on the review site a buy on Amazon button is inserted for me with a short expandable note that compares the Amazon offer with the others already existing on that site. Everything else that would be doable on Amazon should be just dimmed out, as if it won't apply to my context.

This is just scratching the surface, there are so many things that can be connected through the LOD cloud, that you may say that there are




too many possible use cases and no one will create a semantic browser that will cover all of them.

Well, nobody has to. Imagine that the annotations you have in each site are not just about product details, prices and shipping options; let's delegate not only service description to the service providers, but also when those services and options apply. My browser doesn't have to know how to link a birthday in my calendar with Amazon gift wrapping option, Amazon provided the rule that matched a birthday event with its service option, rule which was triggered when I mentioned on the Facebook chat my friend, mention that loaded his info triggering that birthday event. Of course, all of this should happen in the browser and browser alone, your identity and friends won't be disclosed to those remote services as this service composition will be done locally.

This way a semantic browser won't be about browsing semantically related content and services, but it will be about getting things out of the way and about composing meaningful and effective interactions. It will enable a new kind of user experience, a new kind of brand experience.

*A semantic browser won't be about browsing semantically related content and services, but it will be about getting things out of the way and about composing meaningful and effective interactions. It will enable a new kind of user experience, a new kind of brand experience.*

# 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**

# Semantic Web Elevator Pitch

**By David Wood, Vice President of Engineering, Talis Inc.**

Our friend Eric Franzon, VP Community at SemanticWeb.com, recently pleaded for Semantic Web "elevator pitches" to be recorded and placed on YouTube. An elevator pitch is a short story to sell an idea, apocryphally during the duration of an elevator ride. Some, perhaps most, people still consider the Semantic Web difficult to understand so Eric was attempting to gather some ways to explain it.

This article is my attempt at an elevator pitch for the Semantic Web. You can see the original video recording at **http://bit.ly/eDT0Np.**

An elevator, like most of human technology, is a single system designed from the ground up to be just that, a self-contained system. Airplanes, coffee grinders, cars, watches, books and houses are all examples of self-contained systems. The Internet isn't like that at all: it is a network of computer networks, just as the Web is a system of interlinked Web sites. The Semantic Web is a system of interlinked data. What's that good for?

Imagine a New York Times reporter who wants to find out whether an outbreak of a disease could be related to local industrial polluter. Using the Semantic Web, she could find industrial pollution sites and correlate them with reported illnesses, in just a few minutes and without having to submit Freedom of Information Act requests, scour public databases or even make a telephone call.

These things are possible when people or machines publish structured data on the Web and others use that data to answer questions that cut across any individual concern. It is individual concerns that make system self-contained, like a car (to go somewhere) or a coffee grinder (to grind coffee). The Internet, the Web and the Semantic Web are all general purpose, open systems that enable many uses. There is no one, single individual concern for the Internet or the Web. We can't even conceive in advance of all the uses to which they may be put.

In the old dark days before the Internet, researchers would need to track down the


holders of important data. They would have to ask for copies, most often by writing letters. It might take weeks, months or even years to assemble the critical data for a news story or a research article. A researcher needed to coordinate with each data provider to acquire and assemble information of interest.

The Semantic Web enables cooperation without coordination. If the Environmental Protection Agency publishes data about polluters and the Centers for Disease Control publishes data about illnesses by geographic location, the reporter's job is done. She can answer her question by cooperating with those agencies via Semantic Web standards; there is no reason for her to directly coordinate with the providers of the data.

The government agencies published their data using the Resource Description Framework (RDF), a self-describing data format for the Semantic Web. RDF is an international standard of the World Wide Web Consortium, which helps people to agree on ways to describe and publish their data so that others can reuse it.

No one needed to build a special software application for our reporter. It doesn't matter whether the data providers published their information using Flickr, Google Docs or even Craig's List. No one needed to determine in advance what kind of questions she might want to ask. The types of questions possible are limited only by the types of data available.

The same approach can be used to facilitate socialization, work, research and discovery. Anyone can publish data in RDF and anyone who does so makes their data available to the Web community. Even better, you can reuse some of the same relationships used by others and interlink your data to theirs. You can discover more so-called Linked Data at http://linkeddata.org/.

Organizations large and small, government and business and academe have already created Linked Data. The Linked Data on the Web is like a structured version of an encyclopedia. What Wikipedia is to free-form, textual knowledge, the Linked Data collection is to structured data. We can ask questions by querying the Linked Data collection as if it was a huge database.

When data is on the Web, the only thing left to do is to ask meaningful questions. Cooperation without coordination makes that possible.

Other Semantic Web elevator pitches are being collected on http://semanticweb.com. Sandro Hawke of the W3C created the first one and Mark Montgomery of Kyield has also contributed. Have a look or even join in! You can record your own elevator pitch and help explain the Semantic Web to those new to the concept.

# Free Linked Data Open Days

## Schedule 2011

Building on the successful series of free Platform and Linked Data Open Days in 2010 we are delighted to announce an expanded schedule of events for 2011. We are continuing the formula of a relaxed participative format, with opportunities to network lubricated by excellent SPARQL blend coffee and a free lunch.

Talis Linked Data Open Days are an ideal introduction to Linked Data and the Semantic Web either as a broad topic or focused on a particular sector.

### 6th April Linked Data and Health – 76 Portland Place, London

A sharing of understanding, experience, potential, practicalities and benefits of Linked Data in Health and Health Data

### May (date tbc) Linked Data and the Media – London

'The Media' have been at the leading edge of Linked Data - The Guardian, New York Times, BBC and others now make routine use of Linked Data. An open day of sharing of experiences, potential, practicalities and benefits of Linked Data in the media

### June (date tbc) Linked Data in Academia – tbc

Data is core to research, education, and academic administration. There is significant potential for Linked Data . A relaxed open day of sharing of experiences, potential, practicalities and benefits of Linked Data in all aspects of Academia.

### July (date tbc) Linked Data Open Day – Talis Offices, Birmingham

An introduction to Linked Data, its use in the real world, plus an overview of the use of RDF and SPARQL

### 24th July Linked Data and Libraries – The British Library, London

Building on the success of Linked Data and Libraries 2010 an opportunity to share experience with the leaders in the field – The British Library, Talis, and other members of the global library community

For more information and to register for these popular events visit:

**www.talis.com/platform/events**

**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**

# 8th Extended Semantic Web Conference

## May 29th - June 2nd, 2011 | Heraklion, Greece

http://www.eswc2011.org/

## General and Program Chairs

Grigoris Antoniou, University of Crete, GR

Marko Grobelnik, Jozef Stefan Institute, SI
Elena Simperl, Karlsruhe Institute of Technology, DE

## Keynote Speakers

Jim Hendler (Rensselaer Polytechnic Institute)
Abe Hsuan (Irwin & Hsuan LLP)
Prasad Kantamneni (Yahoo!)
Andraž Tori (Zemanta Ltd.)
Chris Welty (IBM Research)

## Open Calls for Workshop Papers

Submission deadline: March 4th, 2011
Notifications: April 1st, 2011
Camera ready version: April 15th, 2011
Workshop Camera-ready: May 2nd, 2011
Workshop Day: May 29th, 2011
or May 30th, 2011

## Workshops

High-Performance Computing for the Semantic Web
Inductive Reasoning and Machine Learning for the Semantic Web
eLearning Approaches for the Linked Data Age
Making Sense of Microposts
Ontology and Semantic Web for Manufacturing
Question Answering over Linked Data
Resource Discovery
Semantic Business Process Management
Semantic Publications
Semantics in Governance and Policy Modelling
User Profile Data in the Social Semantic Web

## Research Paper Topics

Social Web and Web Science
Ontologies
Reasoning
Semantic Data Management
Linked Open Data
Software, Services, Processes and Cloud Computing
Natural Language Processing
Sensor Web
Mobile Web
Semantic Web In-Use
Inductive and probabilistic approaches
Digital Libraries

## AI Mashup Challenge

Submission Deadline: April 1st, 2011

## Registration and Information

Please visit the conference website: http://www.eswc2011.org

brought to you by STI · INTERNATIONAL