Linked Data Horizon Scan

Paul Miller, The Cloud of Data

January 2010

Funding

This work was commissioned, overseen and funded by the Joint Information Systems Committee (JISC).

Declaration of Interest

UK software company, Talis, offers products to the education market that utilise many of the techniques and approaches discussed in this report.

The author is a shareholder and former employee of the company. Every effort has been made to avoid bias and conflict of interest in the preparation of this report.

Licence

Executive Summary

A wealth of valuable data is collected and stored in the systems of complex organisations such as our universities, frequently underutilised for a multitude of reasons from institutional inertia to technological complexity. As budgets contract and competitive pressures increase, the timely and effective exploitation of data is becoming an increasingly important characteristic of the successful organisation; and universities are no exception. From efficiently transparent reporting to data-driven internal decision making and the cost-effective nurturing of new avenues for growth, collaboration or differentiation, there is increasing value in effectively exploiting data to further the institutional mission.

World Wide Web inventor Sir Tim Berners-Lee and others talk compellingly of the value in moving from today's 'Web of Documents' toward a 'Web of Data' in which much of the data we already hold is made available – via the architecture and technologies of the Web itself – for manipulation by computers. Pages on the web meant for reading by people would gain structure, so that while you or I might *read* a postal address off the screen as we do today, software would see the same page and offer to calculate a route to that address, add it to your address book, and more. The research paper stored in your institutional repository would be linked to related papers by the same authors, and placed within context to demonstrate institutional research prowess. The courses offered by your institution would be automatically aggregated with similar courses from elsewhere and made easily accessible to potential students who might never visit your web site or order your prospectus. Relevant data from your institution would be available alongside that from other bodies, powering a range of applications for staff, students, funders, industrial partners and more; the value locked up inside institutional systems would be made available to drive efficiency in today's procedures whilst creating the opportunities for tomorrow's.

This vision of a 'Semantic Web' has been discussed for years, but a combination of political and commercial will, community readiness, technological capability and openly available data has led to a recent leap forward in adoption of one particular aspect of that vision under the banner of Linked Data.

This concept of Linked Data is attracting attention in quarters unfamiliar with the Semantic Web community from which it emerged. Recent announcements from the UK's Prime Minister see the Government join existing implementers as diverse as the BBC, Thomson Reuters, Tesco, Best Buy and Johnson & Johnson.

Four simple principles, or rules, laid down by web inventor Sir Tim Berners-Lee describe the practicalities of Linked Data, and implementers have been quick to apply these in exposing large collections of data for use and reuse, facilitated by the underlying structure of the web itself. In a world in which no single database is comprehensive, the value of being easily able to link related assertions from across diverse data silos is proving compelling.

This report describes Linked Data, and highlights a number of the sectors in which it is already being put to work.

From page 32, a series of recommendations outline ways in which JISC and the wider community might approach the application of Linked Data's rules to good effect.

The core recommendations are reproduced, below.

Web Identifiers

Recommendation 1: review Cabinet Office guidance[1] on the creation of URIs for the UK public sector and W3C guidelines on 'Cool URIs[2].' Draft conformant recommendations for the community.

Recommendation 2: engage the community in identifying a core set of widely used identifiers (probably including existing JACS codes, institutional identifiers, etc) and facilitate or encourage creation of new HTTP URIs in line with the guidance in Recommendation 1. Where necessary, clarify licensing ambiguities to ensure that core identifiers are freely available for exploitation by academic institutions and those building applications on their behalf.

[1] http://www.cabinetoffice.gov.uk/media/308995/public_sector_uri.pdf
[2] http://www.w3.org/TR/cooluris/

Recommendation 3: assess existing infrastructure that members of the community may use in hosting personal profiles, linked to institutional, professional and social network identities as appropriate. Quantify community requirements and identify gaps in provision in order to target additional effort if required.

Data Publishing

Recommendation 4: evaluate the effectiveness of the Office of Public Sector Information (OPSI) Unlocking Service[3], and consider whether a similar approach might be used in helping the community identify data sets to prioritise.

Recommendation 5: evaluate the effectiveness of existing community efforts such as Data Incubator[4], and establish a register of individuals and organisations able to convert data or provide the necessary training. Allocate funding to a number of initial data conversions, prioritising proposals that demonstrate a meeting of data, conversion skills, and an identifiable user community with clearly expressed requirements.

Recommendation 6: undertake work to validate existing data licenses such as those from the Open Data Commons. Actively engage with Government work on data licensing[5]. Disseminate findings via relevant JISC channels e.g. Innovation Support Centres and Advisory Services, and evaluate the feasibility of high level endorsement for an Open Data approach.

[3] http://www.opsi.gov.uk/unlocking-service/OPSIpage.aspx?page=UnlockIndex
[4] http://dataincubator.org/
[5] http://perspectives.opsi.gov.uk/2010/01/licensing-and-datagovuk-launch.html

Recommendation 7: demonstrate the utility of embedding RDFa on institutional web pages by providing funding to add RDFa to course and module descriptions, mandating use of common identifiers such as those offered by JACS. Award funding to demonstrations of added value, such as a UK course finder or a plug-in for a professional body's web site that advertises courses relevant to the profession. Assess the role of XCRI[6] in supporting exposure of course data to the web.

Supporting Measures

Recommendation 8: Identify ways in which the community can *consume* and *contribute to* existing data services.

Recommendation 9: identify a focus for Linked Data activities, perhaps within an existing JISC Innovation Support Centre or Advisory Service. Encourage and assist institutions in exploring Linked Data approaches for themselves.

[6] http://www.xcri.org/

Table of Contents

Introduction

The UK Government's 7 December publication of *Putting the frontline first,*[7] and the 21 January unveiling of the data.gov.uk site[8], mark the latest in a series of significant endorsements for the concept of Linked Data, to which the Prime Minister looks in 'radically opening up publicly held data to promote transparency;'

> "we will aim for the majority of government-published information to be reusable, linked data by June 2011; and we will establish a common licence to reuse data which is interoperable with the internationally recognised Creative Commons model."
>
> (*Putting the frontline first*, p28)

From early beginnings in 2006 as one of Tim Berners-Lee's *Design Issues* notes[9] on the World Wide Web Consortium (W3C) site, Linked Data has recently become a rallying cry for those advocating Government transparency, but it has also found favour with the very different groups in search of new business models for the data-rich enterprise.

Building upon work undertaken on the Semantic Web at W3C[10] and elsewhere, Linked Data takes us some way toward that vision by encouraging and facilitating the exposure of machine-readable data across the Web. Importantly, publication of Linked Data can be achieved without substantial investment in new systems and workflow, whilst quickly creating opportunities for meaningful use and re-use of existing content.

JISC last looked seriously at the Semantic Web in 2005, when Brian Matthews prepared *Semantic Web Technologies*[11] for TechWatch;

> "The Semantic Web is an ambitious vision, first proposed by Tim Berners-Lee, to extend today's Web - imbuing it with a sense of meaning. The articulation of this vision in a now famous article in Scientific American has led to a wide reaching research programme. This programme is resulting in the development of new technologies for describing items of Web-based information and their inter-relationships, but what impact is this development likely to have on Higher and Further Education? This TechWatch report provides an introduction to the Semantic Web - the vision, programme and technologies, but also

[7] http://www.hmg.gov.uk/media/52788/smarter-government-final.pdf
[8] http://data.gov.uk/
[9] http://www.w3.org/DesignIssues/LinkedData.html
[10] http://www.w3.org/2001/sw/
[11] http://www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonscanning/hs0502.aspx

> explains where we currently are in its development and what the likely impact will be on education in areas such as information management and discovery tools, digital libraries, supporting Web-based interaction, and e-learning. It also proposes some realistic timescales for adoption and outlines the current and potential role of the UK F&HE community."

(*Semantic Web Technologies* report abstract, TechWatch site)

Linked Data was, of course, unmentioned, and Matthews' conclusions with respect to the slow-burning Semantic Web's importance to Higher Education proved (perhaps unsurprisingly) muted;

> "The Semantic Web has great potential, and with direct application to the HE and FE sector. However, it has been a long time in development and does require an investment of time, expertise and resources. Nevertheless, the time does seem right to start to think how best to use the simpler applications of the technology.
> So what should HE or FE institutions consider doing now? Institutional libraries should be considering joining collaborations to explore how Semantic Web can best be exploited and investing in training staff, with a view to providing Semantic Web solutions within the next two to three years. Information science professionals and academics working in particular fields should work together to provide the vocabularies and domain ontologies required to support particular fields. Particular communities and research groups could be looking at exploiting the emerging infrastructure to enhance the interaction of their community."

(*Semantic Web Technologies*, pp15-16)

Earlier this year, the *Semantic Technologies in Learning & Teaching*[12] project reported on the current state of the art with respect to provision of tools in the learning and teaching space. Although the project was not initially concerned with Linked Data, the topic is addressed within the final report[13];

> "Analysis of the findings of this report suggests that **building a field of linked open data across UK HE/FE institutions by selectively and securely exposing repositories and institutional data (often data that can already be found on institutions' web pages) can provide significant value** and pave the way for pedagogically meaningful applications powered by application-wide or community-wide agreed ontologies in the future. Encouraging institutions to use linked open data technologies and to document

[12] http://www.semtech.ecs.soton.ac.uk/
[13] http://www.jisc.ac.uk/publications/documents/semantictechnologiesreport.aspx

> successful adoption of semantic technologies is considered of critical importance in this report. HE/FE challenges can be addressed by efficiently linking information across institutions."

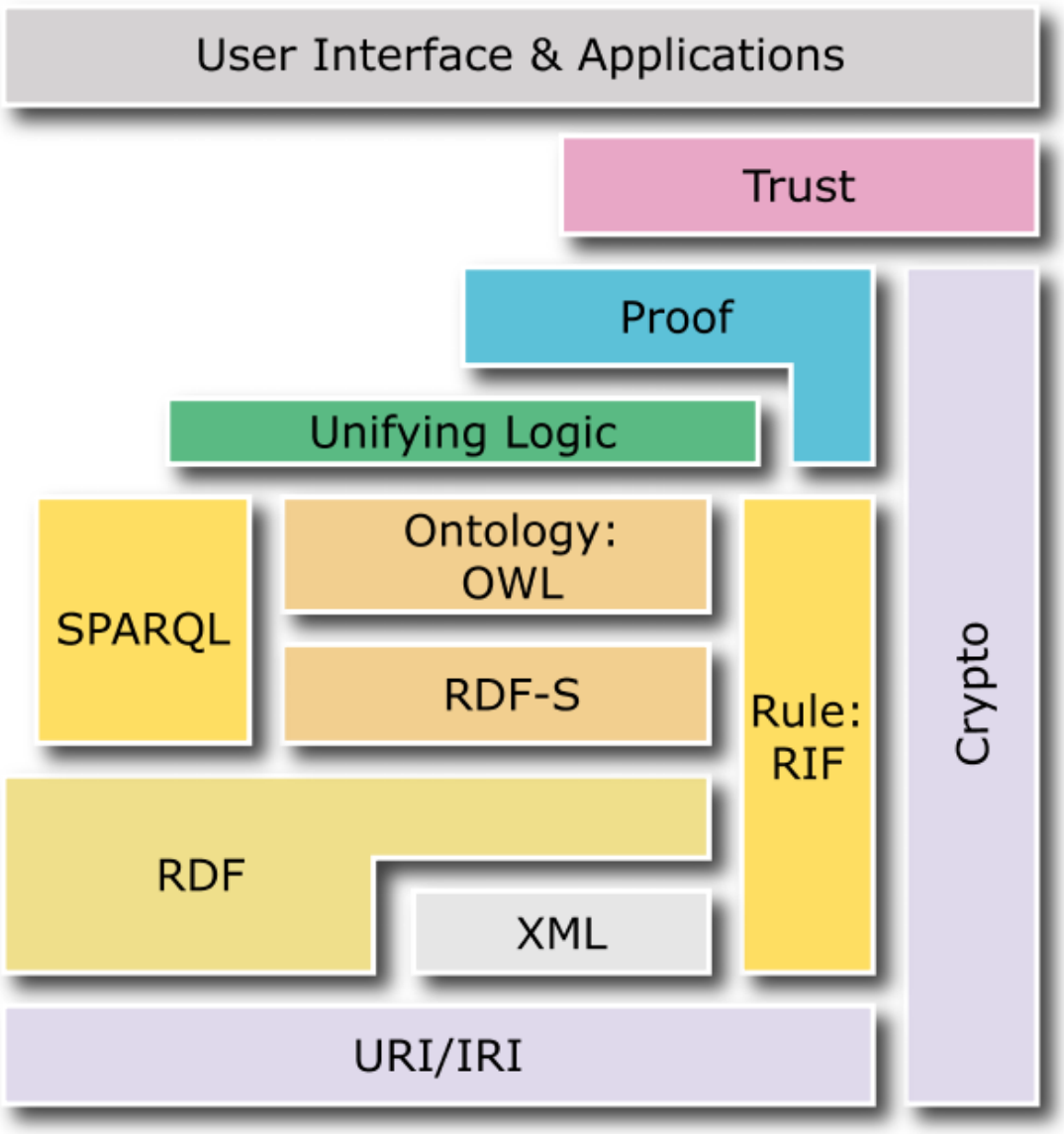
(*JISC SemTech Project Report*, p4. My emphasis)

We begin this latest report – which has been commissioned specifically to explore the opportunities presented to HE by Linked Data – by bringing Matthews' 2005 survey of the Semantic Web up to date, before looking more specifically at the growing interest in Linked Data as a concept both inside Higher Education and beyond.

We conclude by making a series of concrete recommendations to further adoption within the community, identifying areas for future JISC activity as well as pragmatic steps that may be taken in the short term by individual projects, universities, and associated groups.

The Semantic Web

As Matthews notes in his 2005 report, the broad vision of the Semantic Web was essentially laid out for public consumption in a seminal article for *Scientific American* in 2001[14]. Since then, development of the pieces comprising the Semantic Web 'layered architecture[15]' has continued apace.



The Semantic Web Layer Cake[16]

[14] http://www.scientificamerican.com/article.cfm?id=the-semantic-web
[15] http://en.wikipedia.org/wiki/Semantic_Web_Stack
[16] http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/layerCake-4.png

Core components such as RDF[17] were formalised and released as W3C Recommendations over several years, with the 2008 ratification of the SPARQL[18] Query specifications leading Berners-Lee to assert that;

> "I think… we've got all the pieces to be able to go ahead and do pretty much everything… [Y]ou should be able to implement a huge amount of the dream, we should be able to get huge benefits from interoperability using what we've got. So, people are realizing it's time to just go do it."

(Tim Berners-Lee, quoted on ZDNet, http://blogs.zdnet.com/semantic-web/?p=105)

A growing number of companies rely upon semantic technologies within their products, and conferences such as California's Semantic Technology[19] and the smaller European equivalent[20] draw corporate audiences at the same time as companies such as Oracle embed semantic technology within mainstream products and carry them to customers more inclined to attend events on marketing, business intelligence, databases, defence, or other sizeable market areas.

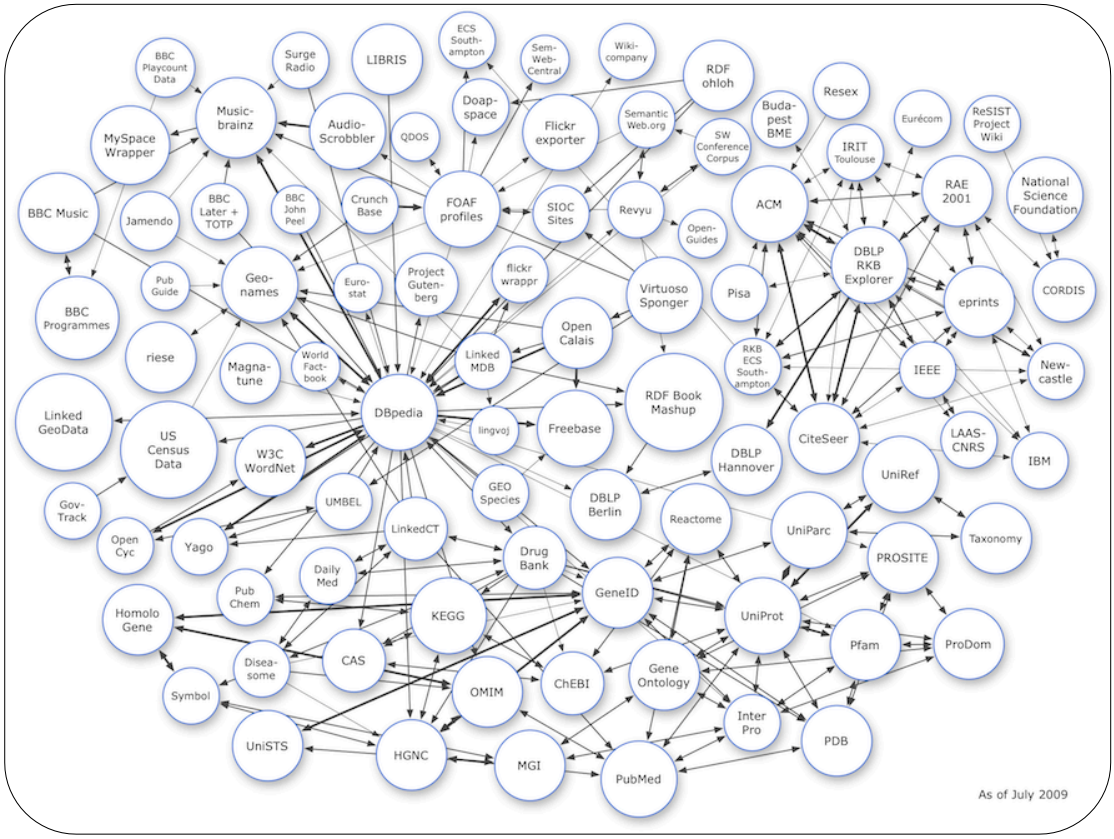
More academic events such as the International[21] and European[22] Semantic Web Conferences continue to draw a research audience, and the European Commission remains a generous funder of Semantic Web projects[23].

The 'Semantic Web' community continues to grow, and encompasses a wide range of perspectives and application areas today. Companies developing solutions for use on highly secure stores of data in the Finance and Defence sectors often appear to have very little in common with those more interested in enriching the structure of the public Web, and communities devoted to the construction of rich ontologies for the expression of nuanced meaning can rub shoulders with those satisfied to leverage the meaning implied in crowd-sourced tags. It is impossible to easily do justice to the scope of these efforts in a short report, and given the specified focus upon the relatively narrow area of 'Linked Data,' it is fortunately unnecessary to try.

---

[17] http://en.wikipedia.org/wiki/Resource_Description_Framework
[18] http://en.wikipedia.org/wiki/Sparql
[19] http://semanticconference.com/
[20] http://www.estc2009.com/
[21] http://iswc2009.semanticweb.org/
[22] http://www.eswc2009.org/
[23] http://blogs.zdnet.com/semantic-web/?p=199

Linked Data



richard.cyganiak.de/2007/10/lod/lod-datasets_2009-07-14.pdf

The concept of Linked Data has been embraced by a particular set of the Semantic Web's enthusiasts and by a growing cohort of potential beneficiaries, predominantly those active in research, media or government. From modest beginnings, Richard Cyganiak's *Linking Open Data Cloud* diagram[24] now represents over 13 billion[25] RDF statements from across a growing network of participating sites. This diagram only scratches the surface, in all likelihood missing a number of poorly publicised resources as well as the related work being done behind the firewalls of organisations such as pharmaceutical giant Johnson & Johnson. The 3,000 datasets currently catalogued on the UK Government's data.gov.uk site are also amongst those not yet represented, having been published since the diagram was last updated.

[24] http://richard.cyganiak.de/2007/10/lod/
[25] http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics

Tim Berners-Lee's Linked Data Principles

As web inventor and W3C Director Sir Tim Berners-Lee notes in his *Design Issues* for Linked Data[26],

> "The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data."

(Linked Data – Design Issues)

This straightforward realisation is expounded in a set of four deceptively simple 'rules' or (as Berners-Lee prefers) 'expectations of behaviour.' Ultimately these lie behind everything that might be described as Linked Data, whether out on the open web for all to see, or locked away in a Computer Science laboratory or behind the firewall at a Pharmaceutical company or Bank.

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things.

(Linked Data – Design Issues)

Whilst the exact wording of these statements has changed slightly since first expressed in 2006[27], and there remains some question[28] as to the strength of the requirement for specific standards, the acronyms mask a simple yet powerful set of behaviours;

1. Name objects and resources, unambiguously;
2. Make use of the structure of the web;
3. Make it easy to discover information about the named object or resource;
4. If you know about related objects or resources, link to them too.

There is a widely held presumption amongst many of Linked Data's most persuasive advocates that the standards (such as RDF[29] for modelling and syntax or SPARQL[30] for querying) referred to by Berners-Lee are prerequisite for sharing or consuming Linked

[26] http://www.w3.org/DesignIssues/LinkedData.html
[27] http://twitter.com/edsu/status/2740552720
[28] http://cloudofdata.com/2009/07/does-linked-data-need-rdf/
[29] http://en.wikipedia.org/wiki/Resource_Description_Framework
[30] http://en.wikipedia.org/wiki/Sparql

Data. Whilst the power of these standards delivers the richest set of capabilities today – with every indication that tool development and the ongoing standardisation process will increase this still further – there is also value in a more permissive reading of Berners-Lee's rules. There is much to gain in embracing the philosophy behind these rules, separately to adopting the standards and specifications required to realise their full potential. Unambiguous identification of resources across the web, easily parsable descriptive information, shared terminologies comprising web-addressable terms, and unambiguous links to related resources deliver real value, as do microformats[31], RDFa[32] markup in web pages and other 'simpler' approaches. Whether this leads towards 'Linked Data' in a formal sense or not perhaps remains unclear, yet may ultimately prove to be unimportant.

As a simple illustration of these principles, let us consider that there are two universities in the English city of York. There is also a York University in Canada. This example is simplistic and there are any number of ways in which people and machines might disambiguate a statement in order to clarify which institution is being referred to, but even so, knowing that the institution in question is 133913 in the Department for Children, Schools and Families' EduBase[33] database or 10007167 in the UK Register of Learning Providers[34] makes for less ambiguity in line with Berners-Lee's first rule.

EduBase also exposes URIs to the web in line with Berners-Lee's second rule, and `www.edubase.gov.uk/establishment/summary.xhtml?urn=133913` refers unambiguously to details of the same institution. The presence of 'summary.xhtml?' in the address may raise issues with respect to persistence as and when the Department makes changes to their software solution, and raises a set of naming issues[35] that have implications far beyond the creation of Linked Data.

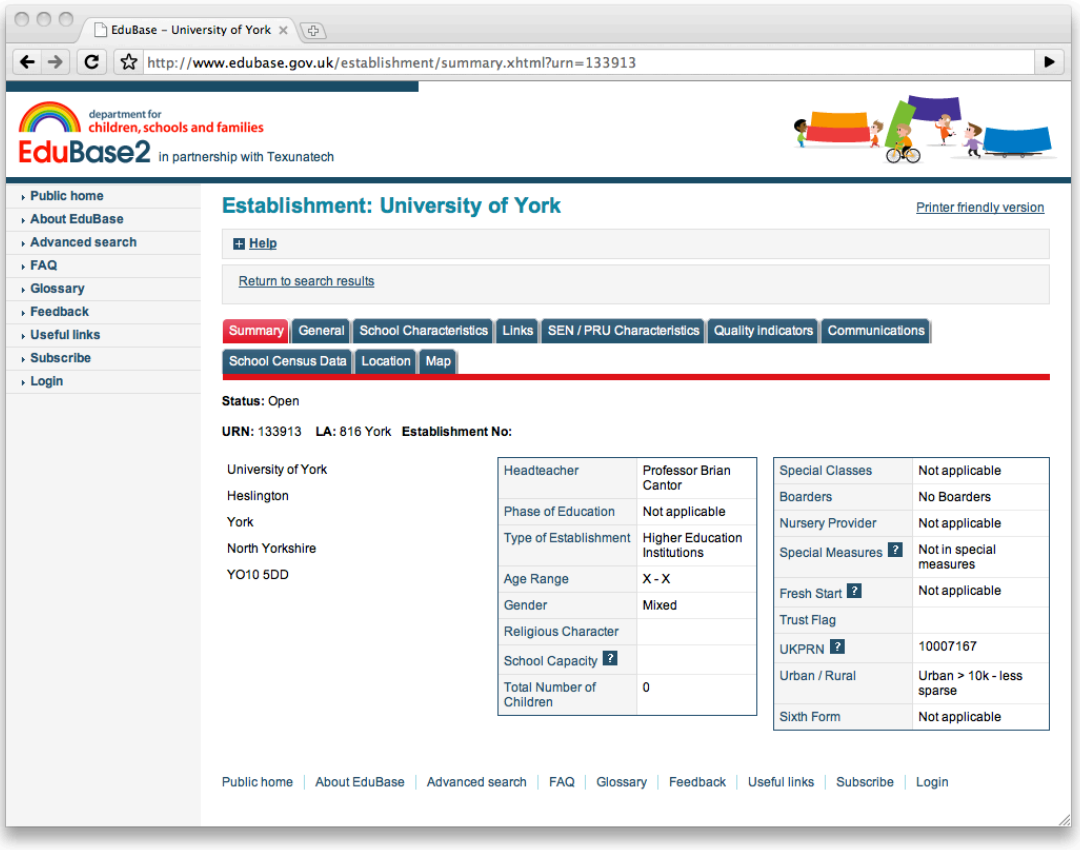
Closely associated with the Linking Open Data Community Project discussed on p18, DBpedia[36] also exposes persistent URIs for the structured information stored in Wikipedia[37]. It is increasingly seen as a reliable means of identifying a wide range of

[31] http://en.wikipedia.org/wiki/Microformat
[32] http://en.wikipedia.org/wiki/Rdfa
[33] http://www.edubase.gov.uk/
[34] http://ukrlp.co.uk/
[35] http://www.w3.org/Provider/Style/URI
[36] http://dbpedia.org/About
[37] http://wikipedia.org/

concepts, including the institution[38] in our example; dbpedia.org/resource/University_of_York. DBpedia has emerged as something of a hub amongst the Linked Data projects, and this trend seems likely to continue.

The value of naming and identification is not, of course, new, although this recent integration with the architecture of the web makes it feasible to consider scalable and sustainable methods of proceeding that encompass both formal naming schemes managed by some responsible authority (ISBNs, DOIs, Learning Provider IDs, etc), more *ad hoc* community efforts such as DBpedia, and even the task or application-specific generation of completely new identifiers as a last resort. There is no requirement to 'boil the ocean,' with massive over-arching schemes that seek to categorise and label *everything* from the outset. Rather, it is increasingly practical to contemplate operating in an environment in which numerous identifiers are assigned to a single resource of interest, and for specific application contexts to rely upon those best suited to their purposes. A particular application might place greatest trust in the institutional identifier assigned by EduBase, but could also include identifiers from DBpedia or the UK RLP; both to support the third and fourth rules, but also to meet the needs of third party applications reliant upon one of these identifiers. There is no need for *every* application to record – and maintain – *every* identifier. Reliance upon the web means that interested applications can traverse the links from one store of data in another, rapidly discovering that 'a' in data store '1' is the same as a resource referred to as 'a' and 'b' in data store '2,' and therefore also the same as 'b' in data store '3.' Until a resource is actually of interest to an application, it may very well go unidentified. By devolving labelling and categorisation to the place and time of need, the larger task becomes more manageable, and the individual identifiers perhaps become more relevant, more timely, and more liable to be actively maintained.

[38] http://dbpedia.org/resource/University_of_York



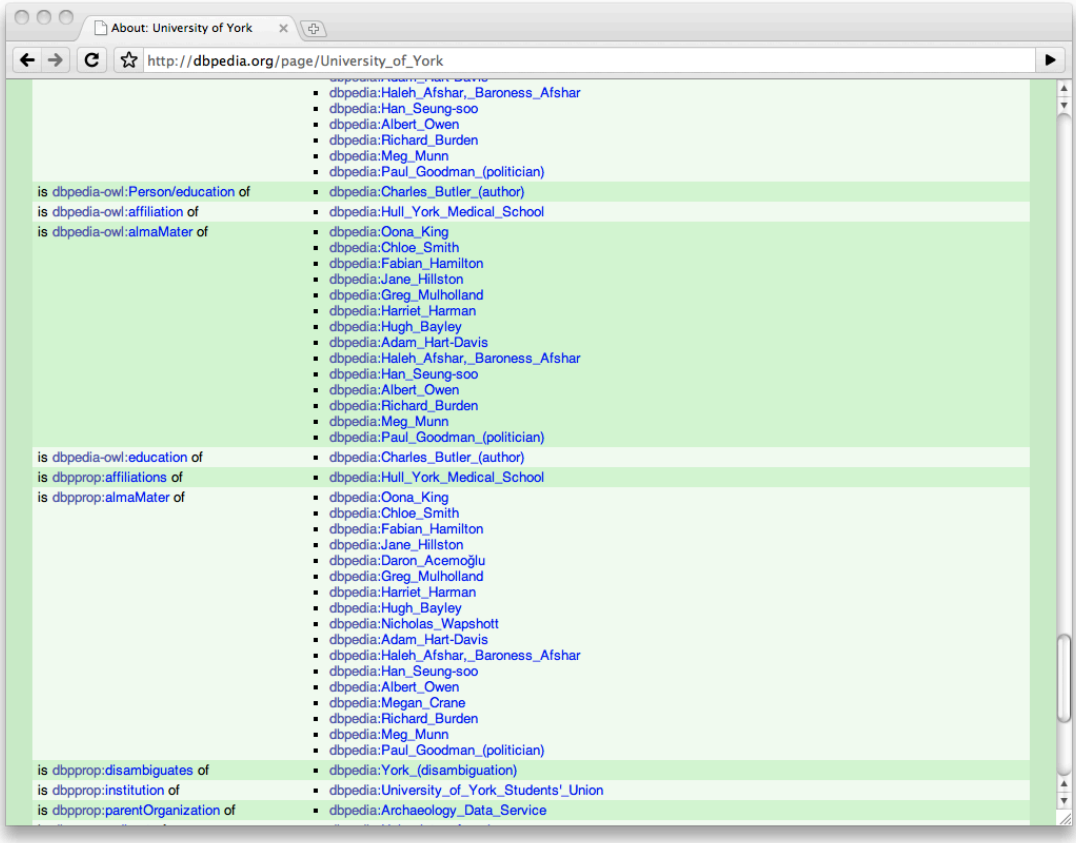
EduBase offers some human-readable 'useful information'

Berners-Lee's third rule calls upon applications to 'provide useful information' when a URI is accessed. EduBase certainly does this, but in a manner that is only really useful for human consumption. To provide useful information in a manner that may be interpreted and acted upon by software, W3C recommendations such as SPARQL[39] and RDF[40] come to the fore, and work has already been done within the UK Government's data.hmg.gov.uk activity to convert EduBase to RDF and to make a SPARQL endpoint[41] available for developers to query[42]. To fully meet Berners-Lee's exhortation to 'provide useful information,' there will be a need to employ content negotiation techniques in order to present different responses to different tools. An undergraduate searching for the University of York is unlikely to welcome being presented with a SPARQL endpoint or an RDF document, and a SPARQL-aware application gathering information about a number of universities will not want human-readable content of the form already delivered from the EduBase web database.

[39] http://www.w3.org/TR/rdf-sparql-query/
[40] http://www.w3.org/RDF/
[41] http://services.data.gov.uk/education/sparql
[42] http://blogs.talis.com/n2/archives/818

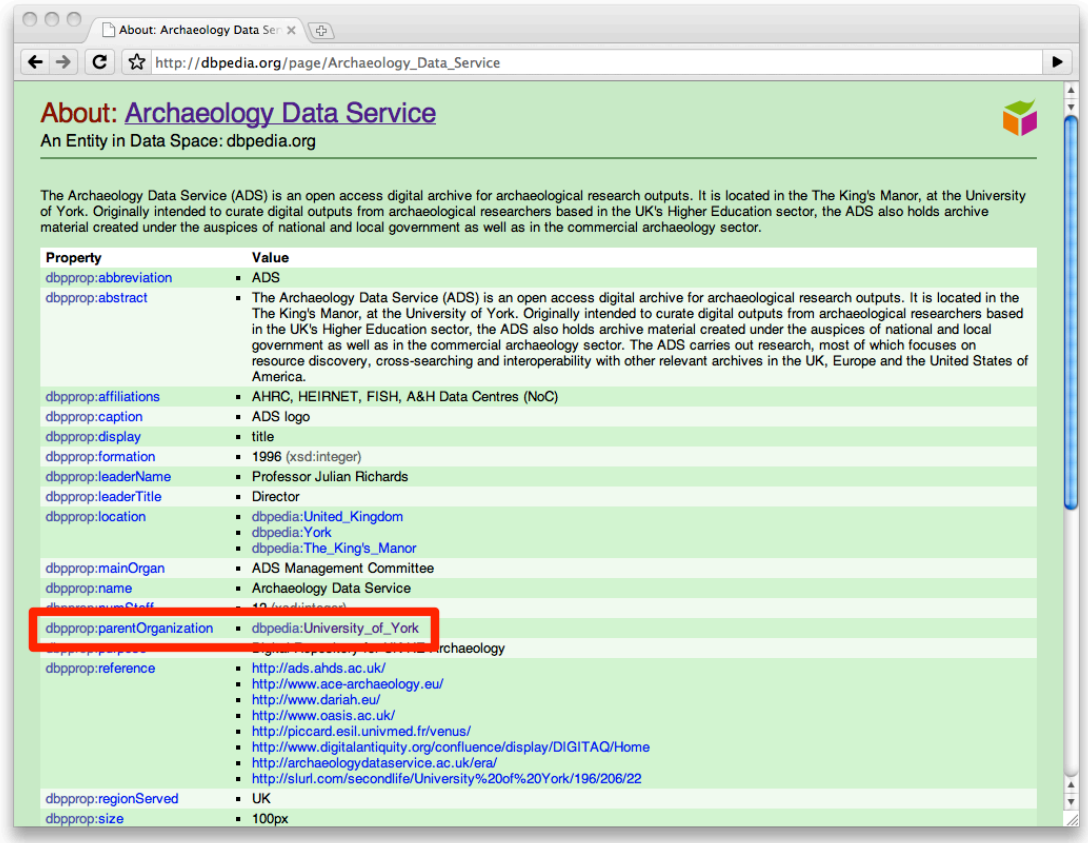
Maintaining wholly separate and unconnected services for humans and for machines makes little sense in the longer term.

Berners-Lee's fourth rule, that resource descriptions should include links to related resources, is well demonstrated by DBpedia.



DBpedia links to famous graduates from the University of York

In this screenshot we can see links to a number of individuals and organisations elsewhere in DBpedia that have declared some relationship to the University. By making it easy – and useful – to declare those links, the web of possible connections grows richer. I might declare myself an *alum* of the University of York, and there may be value in doing so for myself, the University, and third parties interested in either or both of us. Whilst I gain individual value from the relationship, and have an incentive to describe it, the University is more likely to gain value in the aggregate (*all* these people graduated here), and has very little incentive to track an individual such as myself in sufficient detail to declare and maintain the association from their side.



DBpedia page for the Archaeology Data Service, linking back to the University of York as its parent organisation

SWEO Linking Open Data Community Project

The Semantic Web Education and Outreach[43] (SWEO) Interest Group of the World Wide Web Consortium (W3C) was formed in 2006 to;

> "develop strategies and materials to increase awareness among the Web community of the need and benefit for the Semantic Web, and educate the Web community regarding related solutions and technologies."

(SWEO Charter[44])

Concluded in 2008, the Interest Group was responsible for a range of activities including the development of a business case paper[45], the creation of a set of logos[46], and the collection of various business case studies[47]. SWEO also seeded a number of community projects, with the goal of demonstrating the value of Semantic Web

[43] http://www.w3.org/2001/sw/sweo/
[44] http://www.w3.org/2006/07/sweoig-charter.html
[45] http://www.w3.org/2001/sw/sweo/public/BusinessCase
[46] http://www.w3.org/2007/10/sw-logos.html
[47] http://www.w3.org/2001/sw/sweo/public/UseCases/

technologies in the wild. One of these projects was the Linking Open Data Community Project[48], which set out to;

> "extend the Web with a data commons by publishing various open data sets as RDF on the Web and by setting RDF links between data items from different data sources."
>
> (SWEO Linking Open Data Community Project, Project Description)

Participants in this informal activity embraced Berners-Lee's rules and worked to take data they found on the web, convert it to RDF, and begin linking related concepts found in different resources. Perhaps the best known – and most frequently reused – dataset with which the team engaged was DBpedia[49];

> "a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link other data sets on the Web to Wikipedia data. We hope this will make it easier for the amazing amount of information in Wikipedia to be used in new and interesting ways, and that it might inspire new mechanisms for navigating, linking and improving the encyclopaedia itself."
>
> (DBpedia)

DBpedia has become something of a hub for Linked Data projects, with many of them explicitly opting to reuse DBpedia concepts in their own work.

'Linked' Data and 'Open' Data

There is some confusion evident in the way that the terms 'Linked Data,' 'Open Data,' and 'Linked Open Data' are used, often almost interchangeably. SWEO's 'Linking Open Data' project did much to exacerbate this trend, as it grew beyond its original scope to embrace data that were not technically 'Open.'

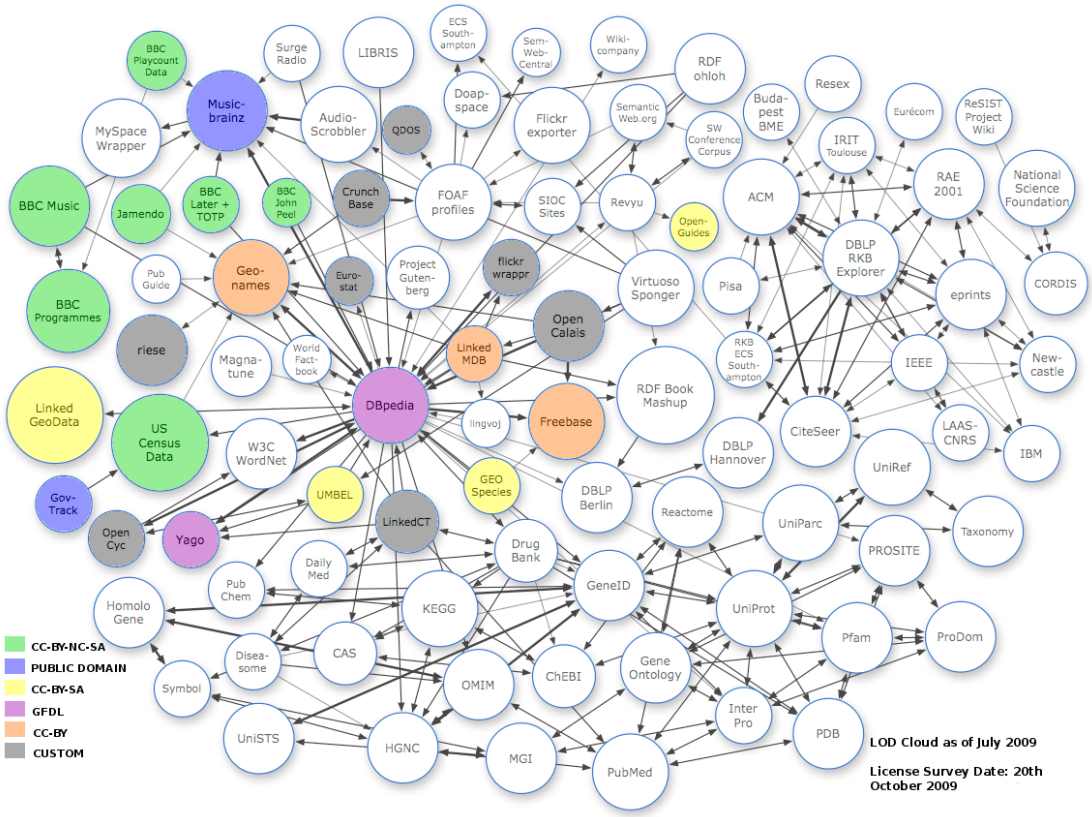
For clarity, 'Linked Data' should normally be presumed to respect Berners-Lee's four rules[50]. 'Open Data' is harder to pin down with precision, but could usefully be considered to cover data respecting the terms of the Open Knowledge Definition[51]. This definition comprises 11 clauses providing detail around the core premise that 'open'

---

[48] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[49] http://dbpedia.org/About
[50] http://www.w3.org/DesignIssues/LinkedData.html
[51] http://opendefinition.org/1.0

data should be freely available online for use and re-use. A number of licenses[52] have been found to be conformant with the Open Knowledge Definition, and should be used where feasible in order to unambiguously assert that data are being made available for re-use.

Linked Data *may* be Open, and Open Data *may* be Linked, but it is equally possible for Linked Data to carry licensing or other restrictions that prevent it being considered Open, or for Open Data to be made available in ways that do not respect all of Berners-Lee's rules for Linking. In order to avoid confusion, the terms 'Linked' and 'Open' should be used with care.



(image by Leigh Dodds, shared on Flickr under Creative Commons license - http://www.flickr.com/photos/ldodds/4043803502/)

A recent exercise by Leigh Dodds of Talis partially illustrates the issue[53]. The now-familiar cloud diagram of Linked Data projects is often referred to as the 'Linked Open Data Cloud' or 'Linking Open Data Cloud,' in deference to the SWEO project from which it evolved. Yet Dodds' analysis clearly illustrates that the majority of datasets carry no explicit open license. It is also questionable whether the licenses chosen in

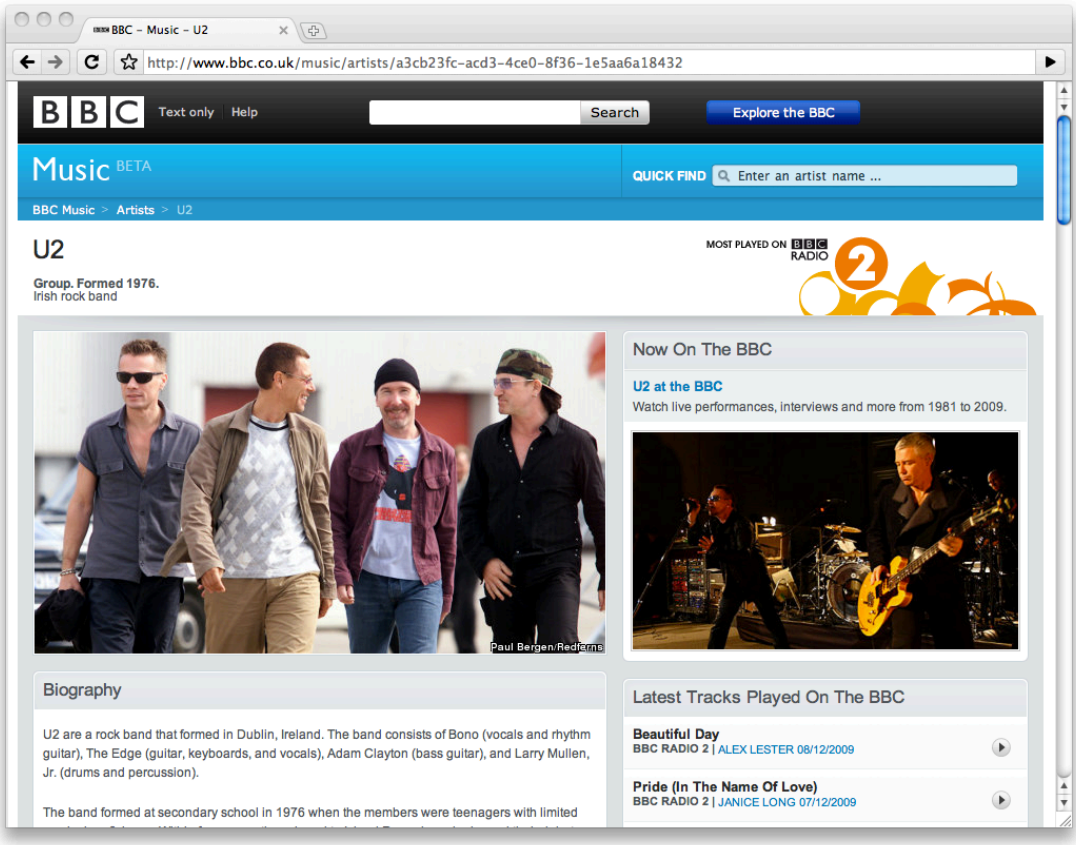
[52] http://opendefinition.org/licenses
[53] http://cloudofdata.com/2009/10/licensing-of-linked-data/

certain cases are appropriate, given the difficulty of applying Copyright-based licenses such as those from Creative Commons to factual data. The work of groups such as the Open Data Commons[54] is relevant in developing licenses appropriate to the use and reuse of data, and should be evaluated for use within Higher Education.

Examples of Success

The early examples of publishing Linked Data tended to be undertaken as experiments, or as part of the work of academics researching the Semantic Web. This work was valuable, and taught the community much about the issues that would need to be overcome. More recently, large organisations have recognised the potential value of Linked Data, and they have begun to publish their own content in this way.

BBC



BBC reuses data from Wikipedia and MusicBrainz to build pages for every artist or band

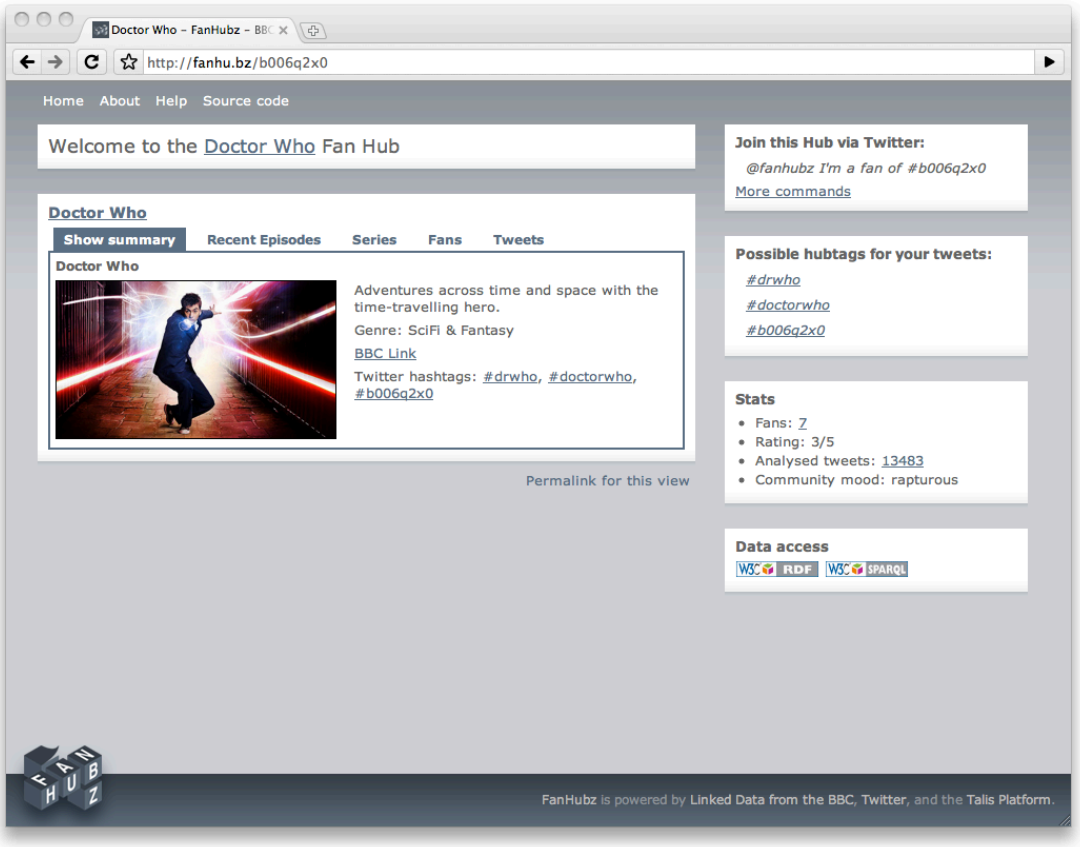
[54] http://www.opendatacommons.org/

The BBC recognises the value of Linked Data[55], and puts these principles to work in a number of recent initiatives, including their Programmes[56] and Music[57] sites. The same approaches are currently being applied to the corporation's Natural History content[58], with discrete identifiers for animals, species, habitats etc.

In each case, concepts (an episode, a series, a performer, a track, an animal) are assigned unique and persistent web URIs (Berners-Lee's first and second rules). Human-readable content is available, as well as representations in RDF, XML, JSON etc that are intended for interpretation by software tools (Berners-Lee's third rule). The Music site re-uses identifiers created by MusicBrainz[59], and displays descriptive content provided by contributors to Wikipedia and MusicBrainz. BBC editorial enhancements are contributed back to MusicBrainz, improving the quality of content available there. Rather than following the more traditional model of specifying, procuring and validating all content in-house, the BBC is actively exploring the opportunities offered by participating in community efforts to build and maintain valuable resources.

The approach being taken by the BBC makes it easier for them to refer in a fine-grained manner to their own content across different properties, and enables them to benefit from externally sourced content such as that on MusicBrainz. The approach also has the added benefit of exposing valuable BBC resources to third party application developers in a manner that makes it straightforward to build products incorporating BBC content. One early example of this is fanhu.bz[60], which builds communities of interest around BBC programmes discussed on Twitter.

[55] http://blogs.talis.com/nodalities/2009/01/building-coherence-at-bbccouk.php
[56] http://www.bbc.co.uk/programmes/developers
[57] http://www.bbc.co.uk/music/developers
[58] http://derivadow.com/2009/07/28/opening-up-the-bbcs-natural-history-archive/
[59] http://musicbrainz.org/
[60] http://fanhu.bz/



Fanhu.bz, displaying data from Twitter and the BBC about *Doctor Who*

New York Times

Earlier this year, the *New York Times* announced its intention[61] to enable access to its thesaurus of more than a million terms describing people, places, organisations, subjects and creative works reported in the paper.

In October, the paper released the first set of data[62]; 5,000 personal names mapped to additional data from Freebase and DBpedia.

As with the BBC examples, data from the *New York Times* is made available in both human readable[63] and machine readable[64] form, simplifying the process of exposing data to browsers visiting a web page and to software aggregating data for some third party application.

[61] http://open.blogs.nytimes.com/2009/06/26/nyt-to-release-thesaurus-and-enter-linked-data-cloud/
[62] http://open.blogs.nytimes.com/2009/10/29/first-5000-tags-released-to-the-linked-data-cloud/
[63] http://data.nytimes.com/N66220017142656459133.html
[64] http://data.nytimes.com/N66220017142656459133.rdf

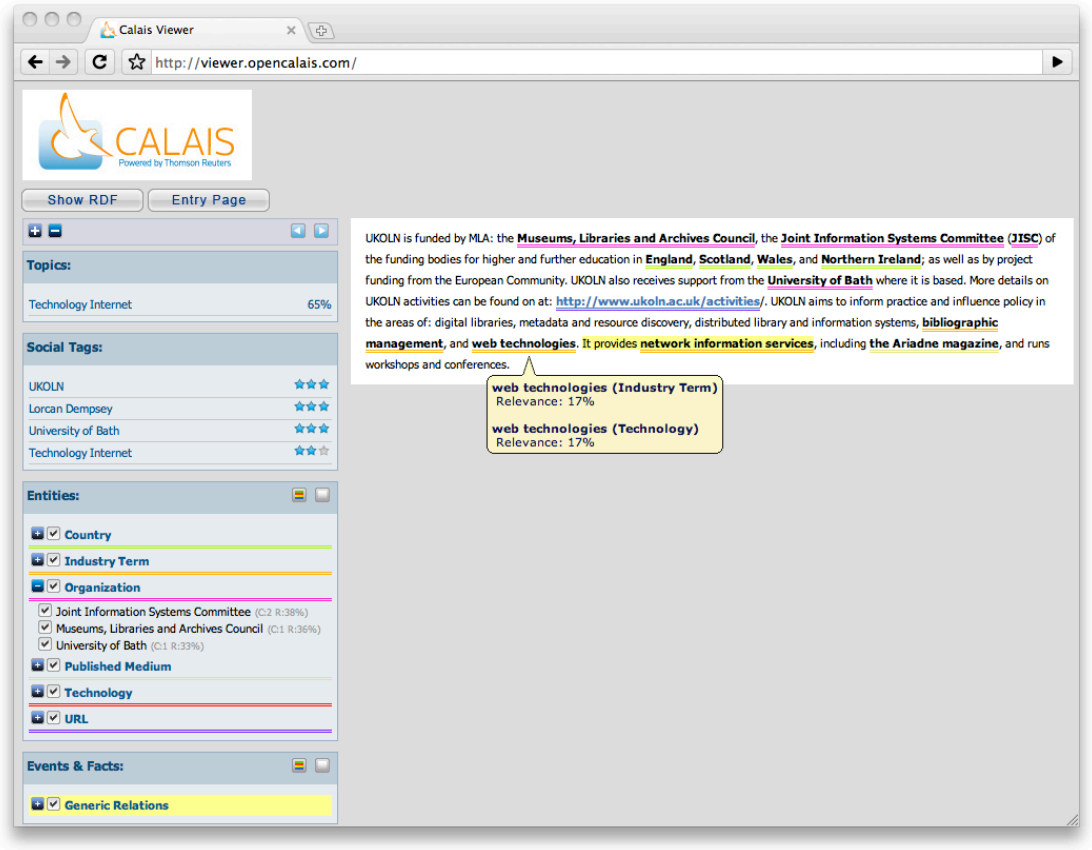
Thomson Reuters

With Open Calais[65], Thomson Reuters offers a free web service that may be used to identify and extract named entities, facts and events from text submitted to it. The service accepts unstructured text submitted in HTML, XML and related formats, and returns a version of the text enriched with additional structure.

Given the heritage of Thomson Reuters, the service tends to be most relevant to business and financial applications, but it succeeds in adding value to a wide range of resource types by extracting meaning from text, adding structure and context, and offering links to a wealth of supporting data from within Thomson Reuters' databases and the third party content of Freebase and others. A passing reference to 'IBM' in text submitted to the Open Calais web service, for example, would be recognised and create the possibility for enrichment with any or all of the additional information known to Thomson Reuters[66] (financial filings, board members, competitors, etc.) or any of the third party services with which Calais shares a common identifier.

In a simple illustration, I copied the first paragraph of UKOLN's 'About' page,[67] stripped out the URLs, and pasted it into the Calais Viewer tool to achieve the result below.

[65] http://opencalais.com/
[66] http://d.opencalais.com/er/company/ralg-tr1r/9e3f6c34-aa6b-3a3b-b221-a07aa7933633.html
[67] http://www.ukoln.ac.uk/about/



A paragraph of text from the UKOLN web site, analysed by Open Calais

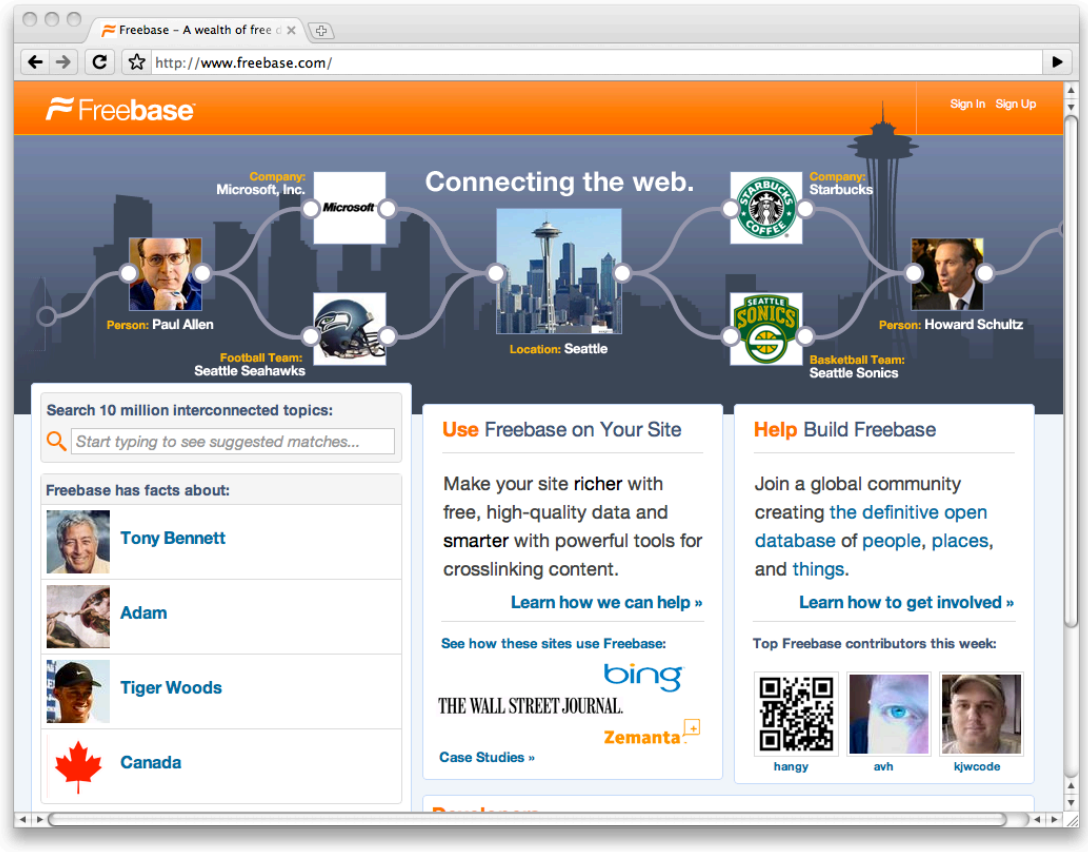
The true potential lies in automated use of the api, rather than manual pasting of demonstration text into a web page, and there is clear scope for individual Higher Education institutions to exploit the connections that a tool such as this identifies.

Freebase

San Francisco-based Freebase[68] is a community-maintained 'free database of the world's information,' backed by significant venture capital.[69] Built upon proprietary database infrastructure, the site offers straightforward tools for expressing rich semantics and structure without directly using the specifications of W3C's Semantic Web stack[70].

[68] http://www.freebase.com/
[69] http://web2innovations.com/money/2008/01/18/massive-second-round-of-funding-for-freebase-42-million/
[70] http://en.wikipedia.org/wiki/Semantic_Web_Stack

Towards the end of 2008 Freebase launched[71] a new RDF service[72] that enabled responses to api calls to be returned in RDF, making Freebase content available to those building Linked Data applications.



Freebase

UK Government

Prime Minister Gordon Brown announced[73] in June that the UK Government intended to make far more of their data easily available online for use and re-use. Sir Tim Berners-Lee was drafted in to help and, far from simply being a figurehead, became actively involved in working with a range of Government departments to make data available online.

Some 3,000 data sets are already available on the Government's data site, data.hmg.gov.uk, and last month's *Putting the frontline first* document reiterates the promise that plenty more will follow. As well as simplifying access to previously 'available' data, there has also been success in changing attitudes to data from

[71] http://blog.freebase.com/2008/10/30/introducing_the_rdf_service/
[72] http://rdf.freebase.com/
[73] http://www.guardian.co.uk/technology/2009/jun/10/berners-lee-downing-street-web-open

Ordnance Survey, the Post Office and other agencies that previously charged significant fees for access.

The UK Government data site, which went public on 21 January 2010

In contrast to the United States' data.gov site, which simply provides access to raw data (Excel spreadsheets, PDF files, and more), the UK is adhering closely to Berners-Lee's Linked Data rules and making data available in formats such as RDF where feasible.

Consumption and Contribution

Linked Data may be consumed from elsewhere to enrich an application, or contributed to the pool for use by others. The norm is, of course, to both consume data provided by others and to contribute your own back to the Commons, but this is certainly not required. A number of commercial organisations consume Linked Data from others using tools such as Open Calais, without giving anything back.

It seems likely that the balance will shift as trust increases, but publication of some compelling case studies would help to illustrate the tangible benefits of reciprocal participation.

The Higher Education Experience

In consulting with the Higher Education community, it is clear that understanding of Linked Data and its implications is not currently widespread. It is worth noting, though, that the techniques and experiences described in this report may well prove to underpin the most cost-effective and sustainable responses to external trends toward transparency and data sharing, such as those implied by *Higher Ambitions*[74]. If the sector is to deliver more robust information about opportunities, outcomes and results, then lightweight and data-based solutions that exploit the existing architecture of the web look more promising than falling back on the traditional methods of procuring yet-more proprietary silos.

Although not directly mentioned in *Putting the frontline first*[75], universities are well placed to embrace the principles outlined here in order to meet existing commitments and prepare new methods to make investments of effort and funding deliver higher returns.

Alongside research into the *practise* of Linked Data from universities such as Southampton, we are beginning to see some early evidence of projects in which Linked Data is put to work as part of a real workflow.

Microsoft's oreChem[76]

> "…is a collaboration between chemistry scholars and information scientists to develop and deploy the infrastructure, services, and applications to enable new models for research and dissemination of scholarly materials in the chemistry community. Although the focus of the project is chemistry, the work is being undertaken with an attention to general cyber infrastructure for eScience, thereby enabling the linkages among disciplines that are required to solve today's key scientific challenges such as global warming. A key aspect of this work, and a core aim of this project, is the design and implementation of an interoperability infrastructure that will allow chemistry scholars to share, reuse, manipulate, and enhance data that are located in repositories, databases, and Web services distributed across the network."
>
> (oreChem project description)

[74] http://www.bis.gov.uk/policies/higher-ambitions
[75] http://www.hmg.gov.uk/frontlinefirst.aspx
[76] http://research.microsoft.com/en-us/projects/orechem/

According to project participant Jim Downing of Cambridge University, the team are extracting named entities from existing Chemistry data and republishing the results as Linked Data for reuse.

Elsewhere, Oxford University's BRII[77]

> "…will enable efficient sharing of research activity information using semantic web technologies. Ontologies and taxonomies will define and describe data objects (eg people, research groups, funding agencies, publications, research 'themes') to forge connections between them and provide web-based services to disseminate and reuse this information in new contexts."
>
> (BRII project description)

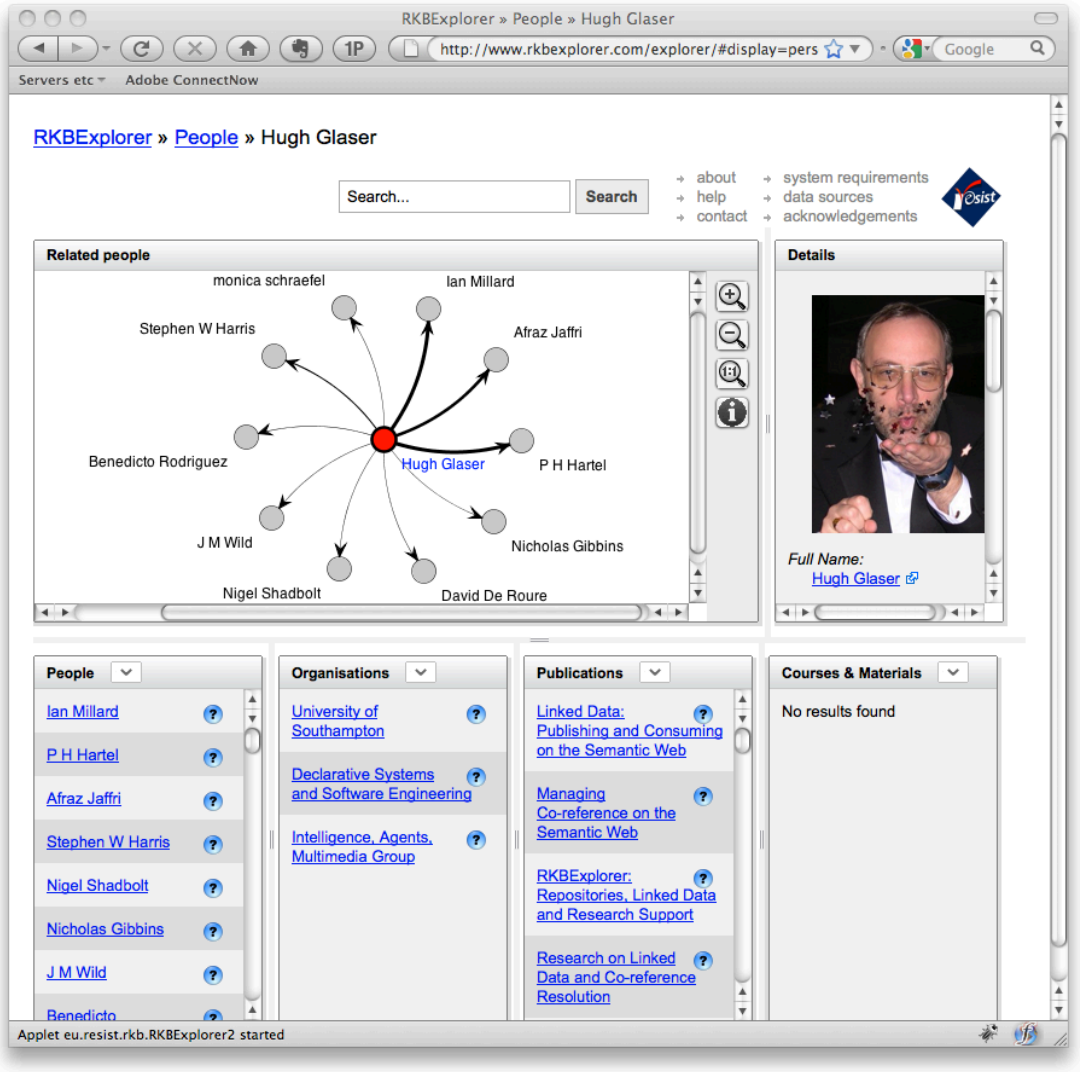
Ben O'Steen (Oxford University Library Services) suggests that use of Linked Data is 'making aggregation easier' as the project team works to combine data from disparate sources across the institution. Key to this are a set of profile pages for academics and departments within the project's pilot group. The profile pages aggregate data previously locked up inside a range of systems, making it far easier to spot mistakes and expose the data to those who might be interested in it. Although similar systems have been built in the past, O'Steen is convinced that the team's current approach results in a flexible system that is more amenable to repurposing. A number of small projects funder under JISC's current programme of Rapid Innovation projects address aspects of the same problem, with Southampton University's dotAC[78] and Heriot-Watt's WattNames[79] amongst those seeking to make progress. dotAC, for example, describes its aim as;

> "to develop a prototype demonstrator that synthesises research data from… research publication metadata from institutional repositories, and research council data and presents it to the end user through an interface that allows them to explore the state of the research landscape in UKHE. "
>
> (dotAC project 'About' page[80])

RKBExplorer,[81] developed at the University of Southampton as part of the European Commission's ReSIST[82] Network of Excellence, illustrates a means of identifying and

[77] http://brii.ouls.ox.ac.uk/
[78] http://www.dotac.info/
[79] http://vreri.googlecode.com/files/Bid15%20WattNames.pdf
[80] http://www.dotac.info/about/

exploring connections between authors, institutions and papers in the scholarly literature.



RKBExplorer

As well as increasingly exploiting data from *within* the sector, universities stand to gain from Government's wider push to open data. Today's EduBase data should be useful in a variety of areas from recruitment and Widening Participation to sociological and educational research, and many of the other datasets will also touch upon a university's diverse academic and administrative interests.

Higher Ambitions and subsequent initiatives from Government are likely to place further transparency demands upon the sector as a whole, increasing the need to find cost-effective and sustainable ways to meet these external demands whilst – ideally – finding

[81] http://www.rkbexplorer.com/
[82] http://www.resist-noe.org/

solutions that also deliver ancillary benefits within the sector. Proactive publication of key metrics might meet future Government demands whilst also addressing institutional and sector objectives, and adoption of a Linked Data-style approach will increase the potential for cost-effective re-use.

Recommendations for Future Work

The growth of Linked Data has been rapid, and early adopters such as the BBC and Thomson Reuters are convinced of the benefits they are seeing. As the scope and scale of these mission-critical applications begin to outstrip the research prototypes that make up much of the core Linked Data community, additional issues such as scalability, provenance, rights and trust become increasingly important. Progress is being made in each area, with commercialisation of increasingly scalable RDF databases[83], ongoing development of more nuanced data licenses[84], and active research into the practical implications of 'trust' and 'provenance' in a richly inter-connected network of resources. These are not problems for JISC – or UK Higher Education – to solve alone, but there is a wealth of expertise in the UK that should continue to participate in these wider international activities.

The recommendations, below, specifically address a number of areas in which UK Higher Education can take action and realise the benefits of doing so. Many of these recommendations are necessarily addressed to JISC and the Higher Education community as a whole, as they are concerned with validating the shared understanding and methodologies that will enable subsequent advances within individual institutions. A growing number of tools already exist to support creation, maintenance and use of this data, but many lack the polish that would be expected before they might usefully be placed before large numbers of non-expert users. By targeting finite resources to specific areas such as those identified below, it is intended to nurture adoption of Linked Data solutions within Higher Education; adoption that will then stimulate the development of end-user tools and applications from both within the community and beyond.

The recommendations fall into two broad categories; those that work behind the scenes laying the groundwork for exposing useful data (broadly addressing Berners-Lee's first two rules from p. 13), and those that identify and publish real data sets onto the Web for use and re-use by all.

---

[83] http://www.tso.co.uk/press/latestnews/archive/2010/triplestore/
[84] http://www.opendatacommons.org/

Web Identifiers

Tim Berners-Lee's Linked Data rules call explicitly for the use of HTTP URIs in naming resources. Although good at creating various schemes of identifiers (such as the JACS codes used to identify courses), the Higher Education sector appears less good at making those identifiers available for effective use over the web.

By exposing existing schemes of identification for institutions, subjects, courses, resources, people and more, myriad opportunities are created for identifying related content, decreasing ambiguity, reducing duplication of effort, and providing lightweight hooks to begin combining data from otherwise incompatible systems.

Scope is created for applications that span subjects or other organisational groupings, as well as those within a single institution that might usefully draw upon authoritative information from further afield.

Recommendation 1: review Cabinet Office guidance[85] on the creation of URIs for the UK public sector and W3C guidelines on 'Cool URIs[86].' Draft conformant recommendations for the community.

There are a number of widely used identification schemes that are currently less accessible than is required if they are to underpin the next generation of services. Common identifiers have, for example, been made 'available' in the form of PDF downloads or searchable via a human-readable query form on a website. *Ad hoc* community efforts such as the Talis-supported[87] Data Incubator[88] and vocab.org[89] provide infrastructure that could be used in converting these core resources to more accessible forms. An earlier version of the JACS codes used to describe undergraduate courses is available via DataIncubator[90], for example, and might usefully be built upon.

[85] http://www.cabinetoffice.gov.uk/media/308995/public_sector_uri.pdf
[86] http://www.w3.org/TR/cooluris/
[87] http://blogs.talis.com/nodalities/2009/05/growing-the-web-of-data-with-data-incubator.php
[88] http://dataincubator.org/
[89] http://vocab.org/
[90] http://jacs.dataincubator.org/

Recommendation 2: engage the community in identifying a core set of widely used identifiers (probably including existing JACS codes, institutional identifiers, etc) and facilitate or encourage creation of new HTTP URIs in line with the guidance in Recommendation 1. Where necessary, clarify licensing ambiguities to ensure that core identifiers are freely available for exploitation by academic institutions and those building applications on their behalf.

As well as identifying institutions, subjects, topics and resources, there is value in unambiguously identifying individuals within universities. The JISC-funded Names Project[91] is exploring the requirements for a service to reliably and uniquely identify individuals and institutions named in the scholarly literature and elsewhere. The project is responding to a requirement that is also being explored at the institutional level through projects such as Oxford's BRII. Web specifications such as FOAF[92] should be of relevance here, and there are opportunities to explore facilitating infrastructure to allow individuals to identify *themselves*, and to link their various professional personas online. By identifying and meeting a clear need, the resulting infrastructure is more likely to be used and kept accurate by its beneficiaries. FOAF Builder[93] from UK-based Garlik illustrates one way in which FOAF might be used to underpin lightweight tools that enable individuals to describe and maintain their own identities, and there may be scope for a similar solution within Higher Education.

From a policy perspective, it makes sense to unambiguously identify and link core groups such as research-active academics, and projects such as RKBExplorer have already illustrated some of the advantages of doing so. Whilst such a model may have wider utility, there are a number of open questions concerning purpose, reach and scope. What (and whom) are these identifiers *for*? Should (and could) they incorporate mappings to existing personas for individuals, for example on social networking sites? Is there a logical relationship with existing institutional solutions such as staff web pages, directories of expertise, contact lists, etc? What does an individual want to expose about themselves, when and where, and how do the requirements of the

[91] http://names.mimas.ac.uk/
[92] http://www.foaf-project.org/

[93] http://foafbuilder.qdos.com/

individual intersect those of their employer? Should an individual carry their identifier from one role or employer to another?

Recommendation 3: assess existing infrastructure that members of the community may use in hosting personal profiles, linked to institutional, professional and social network identities as appropriate. Quantify community requirements and identify gaps in provision in order to target additional effort if required.

Data Publishing

As well as making commonly used identifiers available for easy use and re-use by means of HTTP URIs, there is clearly value in following the examples of both the Linking Open Data Community Project *and* the UK Government in identifying commonly used data sets that the community might benefit from seeing made directly available for use and reuse as RDF. Ready access to rich data from beyond the institution should serve to reduce costs when implementing new systems that require these data, minimise duplication of effort, and create opportunities to make existing applications richer and more accurate by reference to authoritative resources elsewhere. New applications should also be possible that a lack of data previously rendered either too expensive or impractical.

In many cases, existing policy will support making data freely available once identified and (if necessary) converted. In other cases, it may be necessary to explore commercial licensing of resources before they may be used.

Recommendation 4: evaluate the effectiveness of the Office of Public Sector Information (OPSI) Unlocking Service[94], and consider whether a similar approach might be used in helping the community identify data sets to prioritise.

Once identified, there is a role to play in matching data custodians with groups capable of helping convert and/or host the data, and groups of end-users capable of validating whether or not the conversion meets their requirements. The organisations responsible

[94] http://www.opsi.gov.uk/unlocking-service/OPSIpage.aspx?page=UnlockIndex

for maintaining specific resources may wish to make them more widely available, whilst lacking the technical skills to realise their ambition. Equally, many of those with the requisite skills do not directly control relevant data of their own, and both would benefit from some form of clearinghouse that brings them together in order to facilitate a transfer of skills and the building of necessary capacity within the sector.

Recommendation 5: evaluate the effectiveness of existing community efforts such as Data Incubator[95], and establish a register of individuals and organisations able to convert data or provide the necessary training. Allocate funding to a number of initial data conversions, prioritising proposals that demonstrate a meeting of data, conversion skills, and an identifiable user community with clearly expressed requirements.

To permit effective and widespread reuse, data must be explicitly licensed in ways that encourage third party engagement. A growing body of work exists in this area[96], and it is not necessary to reopen the debate. It is, however, necessary to ensure that existing approaches meet the needs of the sector, and to evangelise suitable solutions.

Recommendation 6: undertake work to validate existing data licenses such as those from the Open Data Commons. Actively engage with Government work on data licensing[97]. Disseminate findings via relevant JISC channels e.g. Innovation Support Centres and Advisory Services, and evaluate the feasibility of high level endorsement for an Open Data approach.

The emphasis of Linked Data activities tends to be placed upon sharing large data sets using rich and expressive mechanisms such as RDF, but there is also value in enriching web pages through the addition of structured markup in formats such as RDFa[98] -- especially now that Yahoo! SearchMonkey[99] and Google Rich Snippets[100] explicitly crawl these.

---

[95] http://dataincubator.org/
[96] http://wiki.creativecommons.org/CC0, http://www.opendatacommons.org/
[97] http://perspectives.opsi.gov.uk/2010/01/licensing-and-datagovuk-launch.html
[98] http://en.wikipedia.org/wiki/RDFa
[99] http://developer.yahoo.net/blog/archives/2008/09/searchmonkey_support_for_rdfa_enabled.html

RDFa is particularly useful in adding explicit structure to descriptive content, and might be used in everything from OPAC pages describing a book to the University vacancies site with relative ease. Mark Birbeck, who originally proposed RDFa, recently completed work to add RDFa to job information provided by all central Government departments[101]. The added structure has little impact upon the workflow of those advertising the individual vacancies, and does not affect the look of the page. Behind the scenes, the additional structure enables vacancies from across Government to be aggregated onto a single site, and also raises the visibility of these pages on RDFa-aware search engines. Similar approaches might usefully be taken on university web sites to describe the institutions (on the home page), vacancies, publications, courses, and more.

Although there have been *ad hoc* efforts to make data from specific university systems available in the past, usually on a case-by-case basis, repeating the exercise using RDFa offers a useful introduction to the issues likely to be posed by any wholesale embrace of the Linked Data rules. Certain institutions may well opt to bypass RDFa and move straight to full-blown RDF. For others, this intermediate step may prove more achievable in the short term.

Recommendation 7: demonstrate the utility of embedding RDFa on institutional web pages by providing funding to add RDFa to course and module descriptions, mandating use of common identifiers such as those offered by JACS. Award funding to demonstrations of added value, such as a UK course finder or a plug-in for a professional body's web site that advertises courses relevant to the profession. Assess the role of XCRI[102] in supporting exposure of course data to the web.

[100] http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=99170
[101] http://webbackplane.com/mark-birbeck/blog/2009/04/23/more-rdfa-goodness-from-uk-government-web-sites
[102] http://www.xcri.org/

Supporting Measures

Web-scale data services such as DBpedia, Freebase, Open Calais and others have much to offer in terms of solutions to constructing and scaling core pieces of data infrastructure. These services have also established a strong lead in assigning and maintaining persistent web URIs that the community might usefully seek to reuse, instead of inventing new ones. Equally, universities might take more control of the way in which they are represented by services outside the sector, contributing identifiers and data to these services in such a way that potential users find it easy to link through to university-sourced content.

Recommendation 8: Identify ways in which the community can *consume* and *contribute to* existing data services.

To succeed, any new programme of work requires active championing and evangelism. Tangible short-term outcomes can also be of value in maintaining enthusiasm for a more abstract set of long-term goals.

Recommendation 9: identify a focus for Linked Data activities, perhaps within an existing JISC Innovation Support Centre or Advisory Service. Encourage and assist institutions in exploring Linked Data approaches for themselves.

Acknowledgements

I spoke with a number of individuals during the preparation of this report, including those listed below.

Many thanks to all those who willingly contributed time to share their perspectives. Any errors, omissions or misrepresentations are, of course, my own.

Phil Barker, Heriott-Watt University

Mark Birbeck, webBackplane

Rachel Bruce, JISC

Lorna Campbell, CETIS

Les Carr, University of Southampton

Ken Chad, Ken Chad Consulting

Chris Clarke, Talis

Keith Cole, Mimas

Adam Cooper, CETIS

Hugh Davis, University of Southampton

Leigh Dodds, Talis

Jim Downing, University of Cambridge

David Flanders, JISC

Hugh Glaser, University of Southampton

Lin Goodwin, UCAS

Tony Hirst, Open University

Pete Johnston, Eduserv

Matt Jukes, JISC

David Kay, Sero

Wilbert Kraan, CETIS

Ross MacIntyre, Mimas

Dan Needham, Mimas

Cameron Neylon, STFC

Ben O'Steen, University of Oxford

Joy Palmer, Mimas

Andy Powell, Eduserv

Nadeem Shabir, Talis

Owen Stephens, Open University

Adrian Stevenson, UKOLN

Jane Stevenson, Mimas

Tom Tague, Thomson Reuters

David Tarrant, University of Southampton

Thanassis Tiropanis, University of Southampton

Paul Walk, UKOLN

Jo Walsh, EDINA

I would also like to thank attendees at the SemHE workshop[103] in Nice and participants involved in CETIS' December meeting of their Semantic Technologies Working Group[104].

[103] http://www.semhe.org/
[104] http://jisc.cetis.ac.uk/events/register.php?id=211