

# Linked Data Horizon Scan

Paul Miller, The Cloud of Data

## Funding

The Joint Information Systems Committee (JISC) funded preparation of this report.

## 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



This work is licensed under the *Creative Commons Attribution 2.0 UK: England & Wales Licence*, and any reuse should attribute both the work's author (**Paul Miller, Cloud of Data**) and funder (**Joint Information Systems Committee**.)

To view a copy of this licence, please visit [creativecommons.org/licenses/by/2.0/uk/](http://creativecommons.org/licenses/by/2.0/uk/) or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

# Executive Summary

The 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, 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 29, a series of ten recommendations outline ways in which JISC and the wider community might approach the application of Linked Data's rules to good effect.

These recommendations are reproduced, below.

## Web Identifiers

**Recommendation 1:** review Cabinet Office guidance<sup>1</sup> on the creation of URIs for the UK public sector and W3C guidelines on ‘Cool URIs<sup>2</sup>.’ Draft conformant recommendations for the community.

**Recommendation 2:** identify the core set of widely used identifiers (JACS codes, institutional identifiers, etc) and facilitate 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.

**Recommendation 3:** explore the value of providing infrastructure that members of the community may use in hosting a uniquely identified personal profile, linked to institutional, professional and social network identities as appropriate. Assess whether or not the existing Names infrastructure offers a viable foundation upon which to build.

## Data Publishing

**Recommendation 4:** evaluate the effectiveness of the Office of Public Sector Information (OPSI) Unlocking Service<sup>3</sup>, and consider whether a similar approach might be of value in helping the community identify data sets to prioritise. Take the lead in lobbying for access to frequently requested

---

<sup>1</sup> [http://www.cabinetoffice.gov.uk/media/308995/public\\_sector\\_uri.pdf](http://www.cabinetoffice.gov.uk/media/308995/public_sector_uri.pdf)

<sup>2</sup> <http://www.w3.org/TR/cooluris/>

<sup>3</sup> <http://www.opsi.gov.uk/unlocking-service/OPSIpage.aspx?page=UnlockIndex>

resources, and consider the use of financial incentives to facilitate access where required.

**Recommendation 5:** evaluate the effectiveness of existing community efforts such as Data Incubator<sup>4</sup>, and establish a register of individuals and organisations able to help convert data. Allocate funding to a finite set of data conversions, prioritising proposals that demonstrate a combination of data, conversion skills, and an identifiable user community with clearly expressed requirements.

**Recommendation 6:** fund a focussed piece of work to validate existing data licenses such as those from the Open Data Commons. Actively engage with Government work on data licensing. Use existing JISC Advisory Services to disseminate findings, and evaluate the feasibility of high level endorsement for an Open Data approach.

**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<sup>5</sup> in supporting exposure of course data to the web.

---

<sup>4</sup> <http://dataincubator.org/>

<sup>5</sup> <http://www.xcri.org/>

## Supporting Measures

**Recommendation 8:** explicitly establish internal communications to ensure that activities funded by JISC's Infrastructure & Resources Committee (JIR) are aware of and do not unnecessarily duplicate those started under the auspices of the JISC Learning & Teaching Committee (JLT) and CETIS.

**Recommendation 9:** evaluate the models developed for existing web data services. Consider ways in which the community can *use* existing data infrastructure rather than reinventing it, and *enrich* existing data infrastructure rather than competing with it.

**Recommendation 10:** identify a focus for Linked Data activities, perhaps within an existing JISC Advisory Service. Consider funding attention-raising activities such as a competition for compelling uses of Linked Data within the community.

# Table of Contents

Funding .....	1
Declaration of Interest.....	1
Licence .....	1
Executive Summary .....	2
Web Identifiers .....	3
Data Publishing .....	3
Supporting Measures .....	5
Table of Contents .....	6
Introduction .....	7
The Semantic Web .....	10
Linked Data .....	12
Tim Berners-Lee’s Linked Data Principles .....	13
SWEO Linking Open Data Community Project .....	17
‘Linked’ Data and ‘Open’ Data .....	18
Examples of Success .....	20
BBC .....	20
New York Times .....	22
Thomson Reuters .....	23
Freebase .....	24
UK Government .....	25
Consumption and Contribution .....	26
The Higher Education Experience .....	27
Recommendations for Future Work .....	29
Web Identifiers .....	29
Data Publishing .....	31
Supporting Measures .....	33
Acknowledgements .....	35

# Introduction

The UK Government's 7 December publication of *Putting the frontline first*<sup>6</sup> marks 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<sup>7</sup> 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<sup>8</sup> 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*<sup>9</sup> 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

<sup>6</sup> <http://www.hmg.gov.uk/media/52788/smarter-government-final.pdf>

<sup>7</sup> <http://www.w3.org/DesignIssues/LinkedData.html>

<sup>8</sup> <http://www.w3.org/2001/sw/>

<sup>9</sup> <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 CETIS-funded *Semantic Technologies in Learning & Teaching*<sup>10</sup> 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<sup>11</sup>;

“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

<sup>10</sup> <http://www.semtech.ecs.soton.ac.uk/>

<sup>11</sup> <http://www.jisc.ac.uk/publications/documents/semantictchnologiesreport.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)*

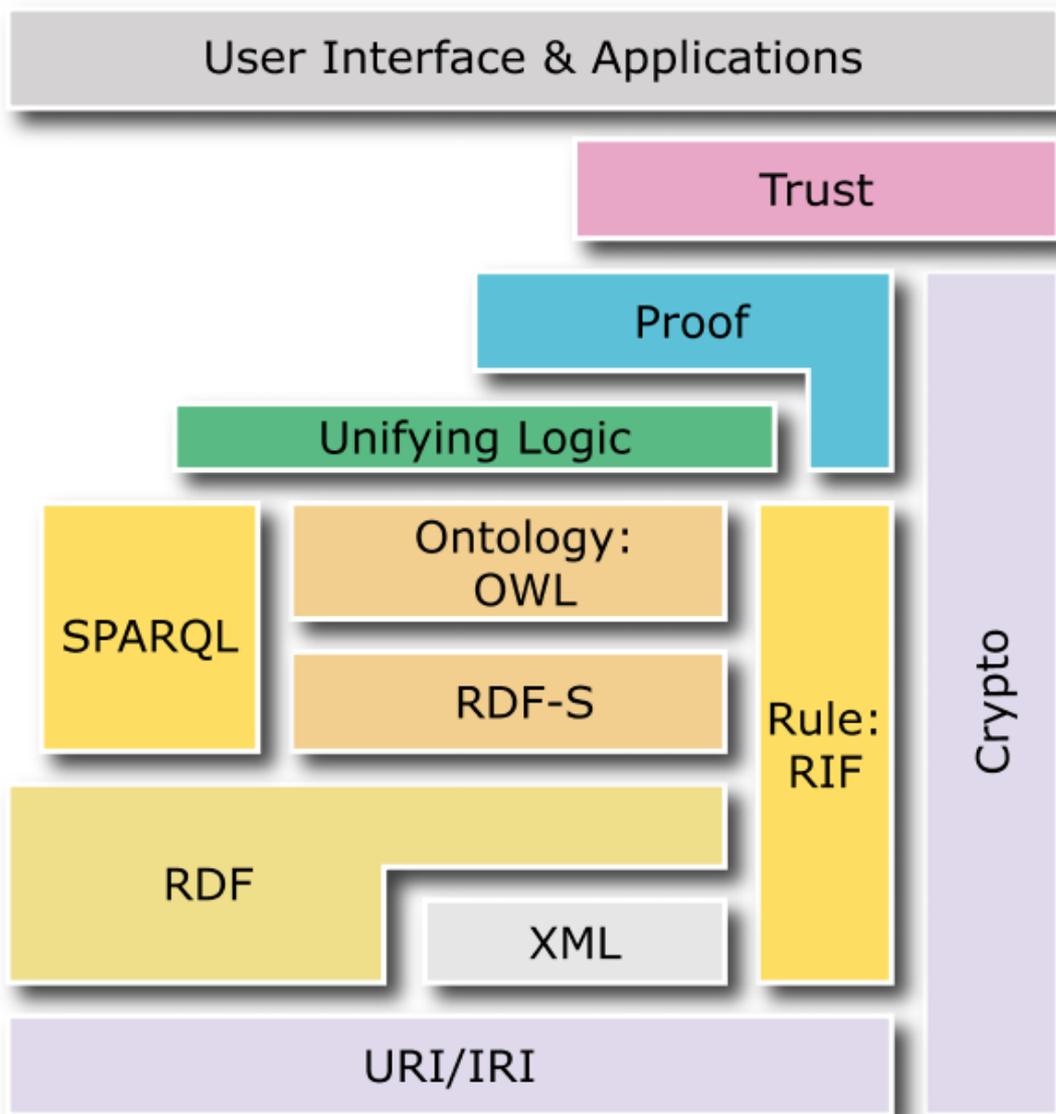
We begin this report by bringing Matthews' 2005 survey of the Semantic Web up to date, before looking more specifically at the growing interest in Linked Data as a concept both inside Higher Education and beyond.

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.

DRAFT

# 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<sup>12</sup>. Since then, development of the pieces comprising the Semantic Web 'layered architecture'<sup>13</sup> has continued apace.



The Semantic Web Layer Cake<sup>14</sup>

<sup>12</sup> <http://www.scientificamerican.com/article.cfm?id=the-semantic-web>

<sup>13</sup> [http://en.wikipedia.org/wiki/Semantic\\_Web\\_Stack](http://en.wikipedia.org/wiki/Semantic_Web_Stack)

<sup>14</sup> <http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/layerCake-4.png>

Core components such as RDF<sup>15</sup> were formalised and released as W3C Recommendations over several years, with the 2008 ratification of the SPARQL<sup>16</sup> 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<sup>17</sup> and the smaller European equivalent<sup>18</sup> 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<sup>19</sup> and European<sup>20</sup> Semantic Web Conferences continue to draw a research audience, and the European Commission remains a generous funder of Semantic Web projects<sup>21</sup>.

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.

---

<sup>15</sup> [http://en.wikipedia.org/wiki/Resource\\_Description\\_Framework](http://en.wikipedia.org/wiki/Resource_Description_Framework)

<sup>16</sup> <http://en.wikipedia.org/wiki/Sparql>

<sup>17</sup> <http://semanticconference.com/>

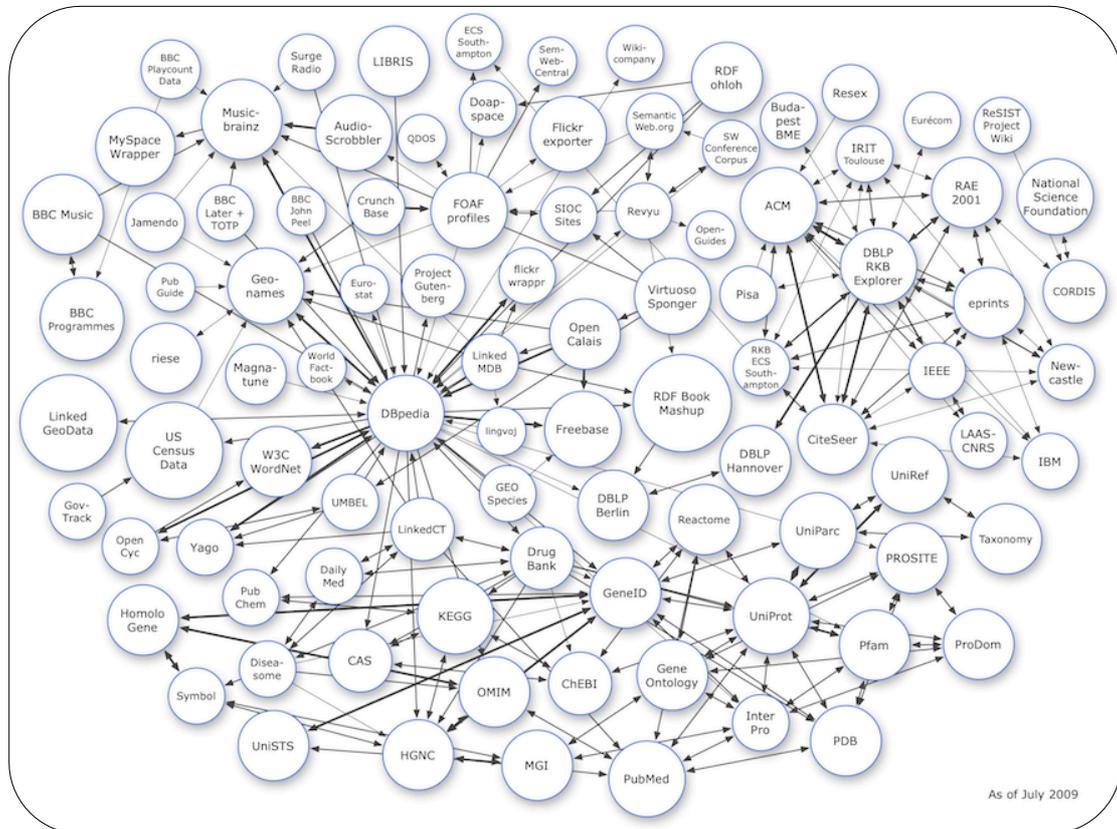
<sup>18</sup> <http://www.estc2009.com/>

<sup>19</sup> <http://iswc2009.semanticweb.org/>

<sup>20</sup> <http://www.eswc2009.org/>

<sup>21</sup> <http://blogs.zdnet.com/semantic-web/?p=199>

# Linked Data



[richard.cyaniak.de/2007/10/lod/lod-datasets\\_2009-07-14.pdf](http://richard.cyaniak.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 Cyaniak's *Linking Open Data Cloud* diagram<sup>22</sup> now represents over 13 billion<sup>23</sup> 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.

<sup>22</sup> <http://richard.cyaniak.de/2007/10/lod/>

<sup>23</sup> <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*<sup>24</sup>,

“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<sup>25</sup>, and there remains some doubt<sup>26</sup> 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 are, for example, 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

<sup>24</sup> <http://www.w3.org/DesignIssues/LinkedData.html>

<sup>25</sup> <http://twitter.com/edsu/status/2740552720>

<sup>26</sup> <http://cloudofdata.com/2009/07/does-linked-data-need-rdf/>

133913 in the Department for Children, Schools and Families' EduBase<sup>27</sup> database or 10007167 in the UK Register of Learning Providers<sup>28</sup> 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](http://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<sup>29</sup> that have implications far beyond the creation of Linked Data.

Closely associated with the Linking Open Data Community Project discussed on p17, DBpedia<sup>30</sup> also exposes persistent URIs for the structured information stored in Wikipedia<sup>31</sup>. It is increasingly seen as a reliable means of identifying a wide range of concepts, including the institution<sup>32</sup> in our example; [dbpedia.org/resource/University\\_of\\_York](http://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 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. It is feasible to contemplate operating in an environment in which numerous identifiers exist for a single resource, 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

---

<sup>27</sup> <http://www.edubase.gov.uk/>

<sup>28</sup> <http://ukrlp.co.uk/>

<sup>29</sup> <http://www.w3.org/Provider/Style/URI>

<sup>30</sup> <http://dbpedia.org/About>

<sup>31</sup> <http://wikipedia.org/>

<sup>32</sup> [http://dbpedia.org/resource/University\\_of\\_York](http://dbpedia.org/resource/University_of_York)

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.'

The screenshot shows the EduBase website interface for the University of York. The page title is "Establishment: University of York". The URL is <http://www.edubase.gov.uk/establishment/summary.xhtml?urn=133913>. The page features a navigation menu on the left with options like "Public home", "About EduBase", "Advanced search", "FAQ", "Glossary", "Feedback", "Useful links", "Subscribe", and "Login". The main content area includes a "Help" button, a "Return to search results" link, and a set of tabs for "Summary", "General", "School Characteristics", "Links", "SEN / PRU Characteristics", "Quality Indicators", and "Communications". Below these tabs are sub-tabs for "School Census Data", "Location", and "Map". The "Summary" tab is active, displaying the following information:

**Status:** Open  
**URN:** 133913 **LA:** 816 York **Establishment No.:**

University of York	Headteacher	Professor Brian Cantor	Special Classes	Not applicable
Heslington	Phase of Education	Not applicable	Boarders	No Boarders
York	Type of Establishment	Higher Education Institutions	Nursery Provider	Not applicable
North Yorkshire	Age Range	X - X	Special Measures ?	Not in special measures
YO10 5DD	Gender	Mixed	Fresh Start ?	Not applicable
	Religious Character		Trust Flag	
	School Capacity ?		UKPRN ?	10007167
	Total Number of Children	0	Urban / Rural	Urban > 10k - less sparse
			Sixth Form	Not applicable

At the bottom of the page, there is a footer with links: "Public home | About EduBase | Advanced search | FAQ | Glossary | Feedback | Useful links | Subscribe | Login".

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<sup>33</sup> and RDF<sup>34</sup> 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<sup>35</sup> available for developers to query<sup>36</sup>. 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

<sup>33</sup> <http://www.w3.org/TR/rdf-sparql-query/>

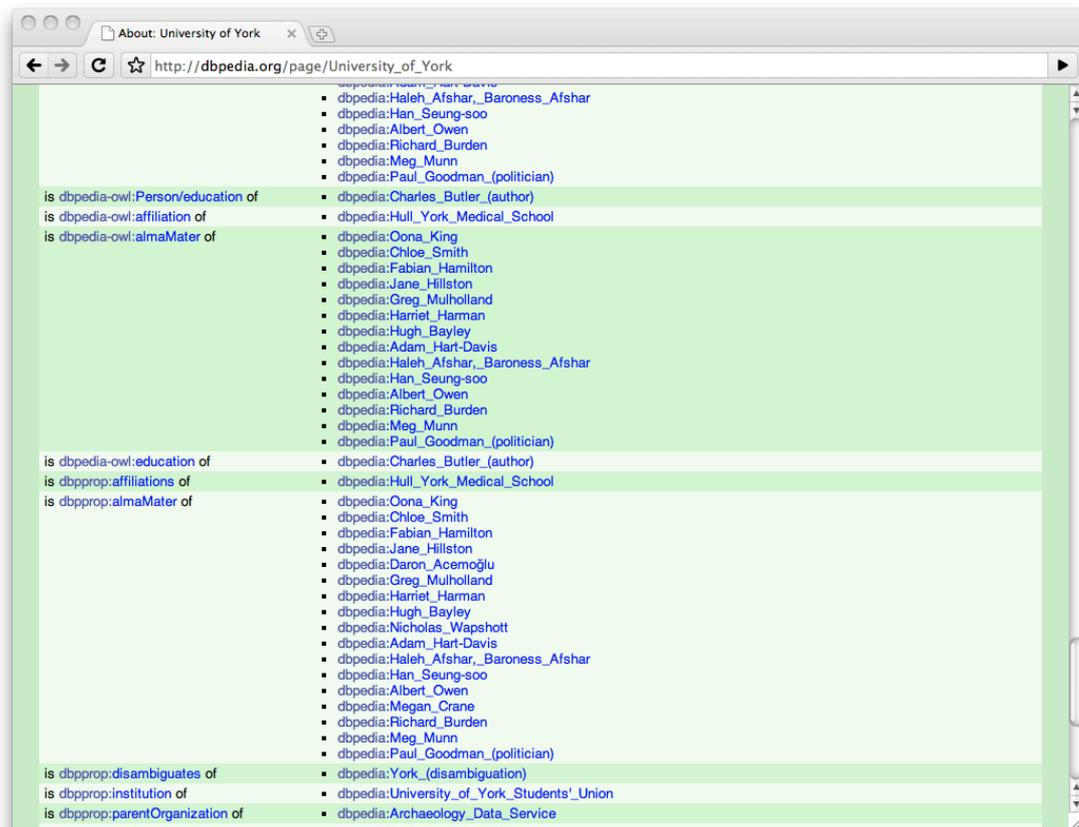
<sup>34</sup> <http://www.w3.org/RDF/>

<sup>35</sup> <http://services.data.gov.uk/education/sparql>

<sup>36</sup> <http://blogs.talis.com/n2/archives/818>

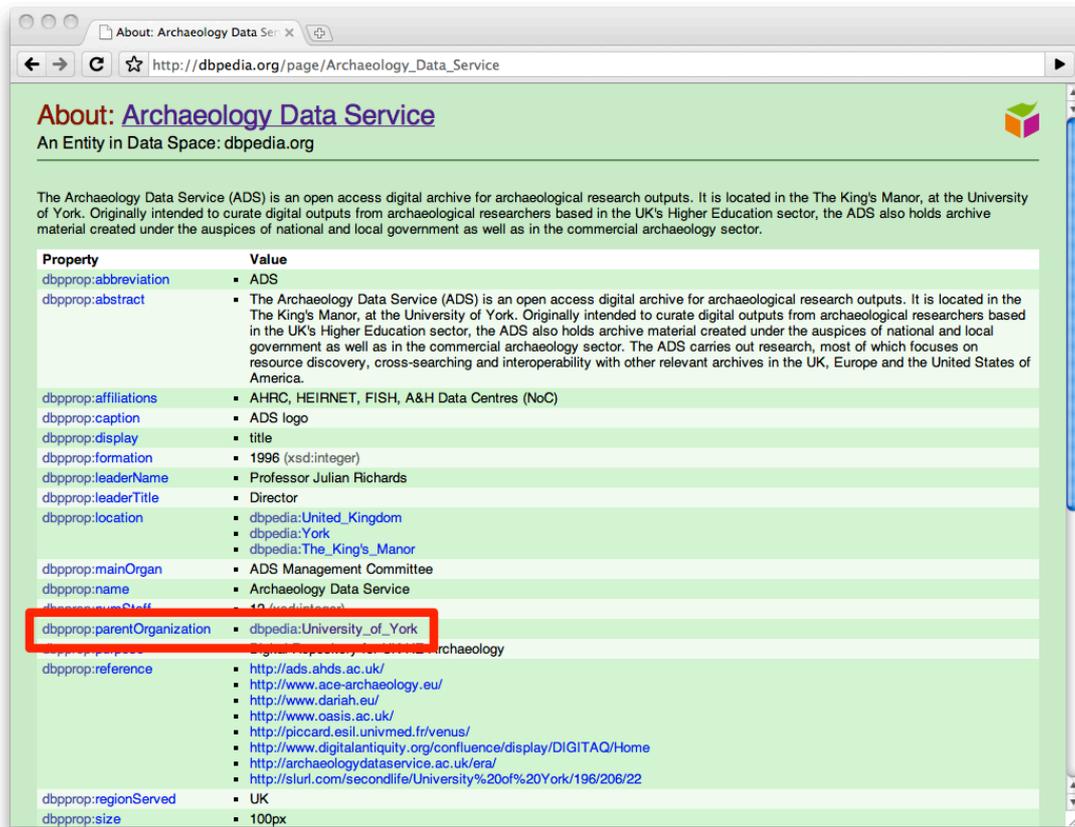
application gathering information about a number of universities will not want human-readable content of the form already delivered from the EduBase web database. Maintaining wholly separate and unconnected services for humans and for machines makes little sense in the longer term.

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<sup>37</sup> (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<sup>38</sup>)

Concluded in 2008, the Interest Group was responsible for a range of activities including the development of a business case paper<sup>39</sup>, the creation of a set of logos<sup>40</sup>, and the collection of various business case studies<sup>41</sup>. SWEO also seeded a number of community projects, with the goal of demonstrating the value of Semantic Web

<sup>37</sup> <http://www.w3.org/2001/sw/sweo/>

<sup>38</sup> <http://www.w3.org/2006/07/sweoig-charter.html>

<sup>39</sup> <http://www.w3.org/2001/sw/sweo/public/BusinessCase>

<sup>40</sup> <http://www.w3.org/2007/10/sw-logos.html>

<sup>41</sup> <http://www.w3.org/2001/sw/sweo/public/UseCases/>

technologies in the wild. One of these projects was the Linking Open Data Community Project<sup>42</sup>, 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<sup>43</sup>;

“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 in the cloud of 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<sup>44</sup>. ‘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<sup>45</sup>. This definition comprises 11 clauses providing detail around the core premise that ‘open’

<sup>42</sup> <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

<sup>43</sup> <http://dbpedia.org/About>

<sup>44</sup> <http://www.w3.org/DesignIssues/LinkedData.html>

<sup>45</sup> <http://opendefinition.org/1.0>

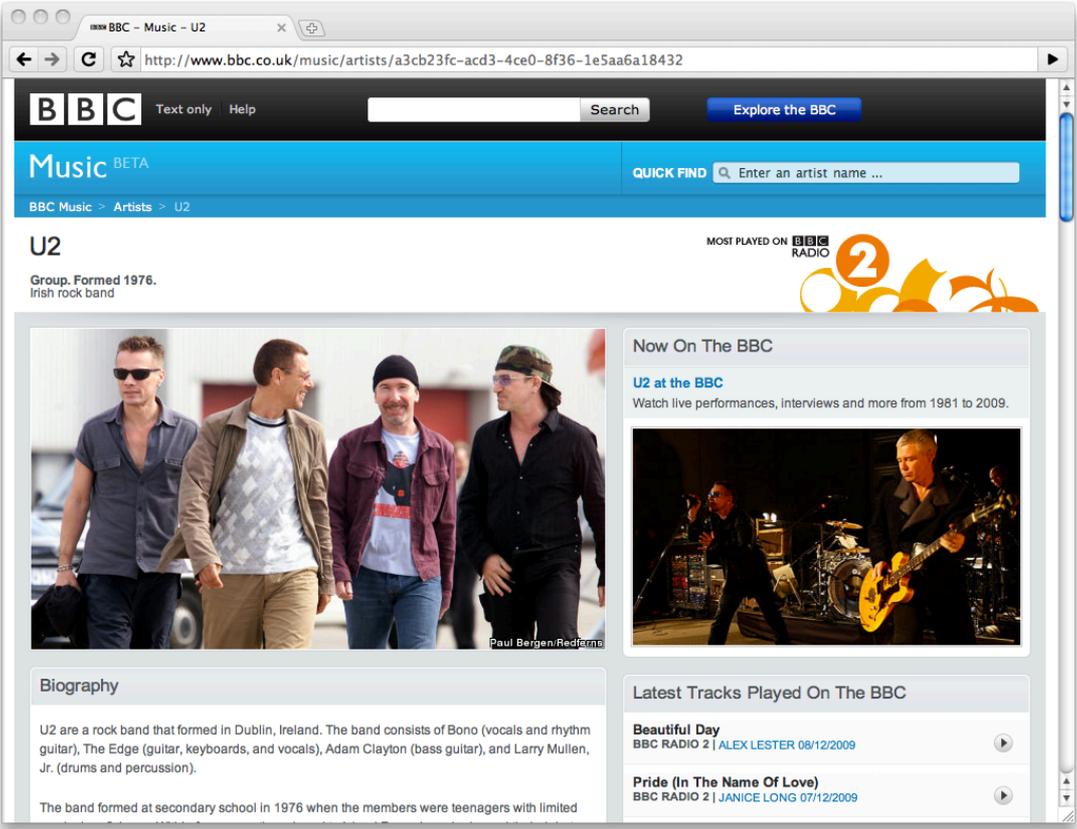


certain cases are appropriate, given the difficulty of applying Copyright-based licenses such as those from Creative Commons to factual data. The work of groups such as the Open Data Commons<sup>48</sup> is relevant in developing licenses appropriate to the use and reuse of data, and should be evaluated 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



The screenshot shows the BBC Music website for the artist U2. The page features a navigation bar with the BBC logo, a search bar, and a 'QUICK FIND' section. The main content area includes a large photo of the band U2, a biography section, and a 'Now On The BBC' section with a video player. The biography section contains the following text:

**Biography**

U2 are a rock band that formed in Dublin, Ireland. The band consists of Bono (vocals and rhythm guitar), The Edge (guitar, keyboards, and vocals), Adam Clayton (bass guitar), and Larry Mullen, Jr. (drums and percussion).

The band formed at secondary school in 1976 when the members were teenagers with limited

The 'Now On The BBC' section includes a video player for 'U2 at the BBC' and a list of 'Latest Tracks Played On The BBC':

- Beautiful Day**  
BBC RADIO 2 | ALEX LESTER 08/12/2009
- Pride (In The Name Of Love)**  
BBC RADIO 2 | JANICE LONG 07/12/2009

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

<sup>48</sup> <http://www.opendatacommons.org/>

The BBC recognises the value of Linked Data<sup>49</sup>, and puts these principles to work in a number of recent initiatives, including their Programmes<sup>50</sup> and Music<sup>51</sup> sites. The same approaches are currently being applied to the corporation's Natural History content<sup>52</sup>, 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<sup>53</sup>, 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<sup>54</sup>, which builds communities of interest around BBC programmes discussed on Twitter.

---

<sup>49</sup> <http://blogs.talis.com/nodalities/2009/01/building-coherence-at-bbccouk.php>

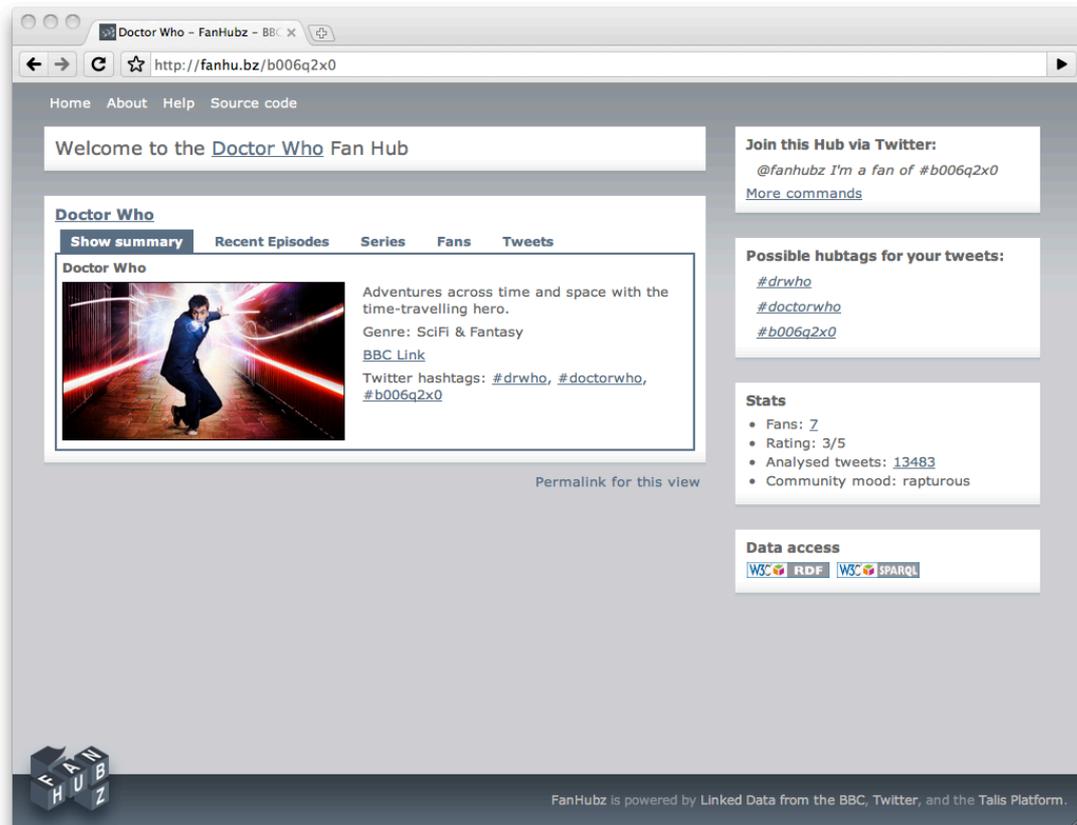
<sup>50</sup> <http://www.bbc.co.uk/programmes/developers>

<sup>51</sup> <http://www.bbc.co.uk/music/developers>

<sup>52</sup> <http://derivadow.com/2009/07/28/opening-up-the-bbcs-natural-history-archive/>

<sup>53</sup> <http://musicbrainz.org/>

<sup>54</sup> <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<sup>55</sup> 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<sup>56</sup>; 5,000 personal names mapped to additional data from Freebase and DBpedia.

As with the BBC examples, data from the *Times* is made available in both human readable<sup>57</sup> and machine readable<sup>58</sup> form, simplifying the process of exposing data to browsers visiting a web page and to software aggregating data for some third party application.

<sup>55</sup> <http://open.blogs.nytimes.com/2009/06/26/nyt-to-release-thesaurus-and-enter-linked-data-cloud/>

<sup>56</sup> <http://open.blogs.nytimes.com/2009/10/29/first-5000-tags-released-to-the-linked-data-cloud/>

<sup>57</sup> <http://data.nytimes.com/N66220017142656459133.html>

<sup>58</sup> <http://data.nytimes.com/N66220017142656459133.rdf>

### *Thomson Reuters*

With Open Calais<sup>59</sup>, 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<sup>60</sup> (financial filings, board members, competitors, etc.) or any of the third party services with which Calais shares a common identifier.

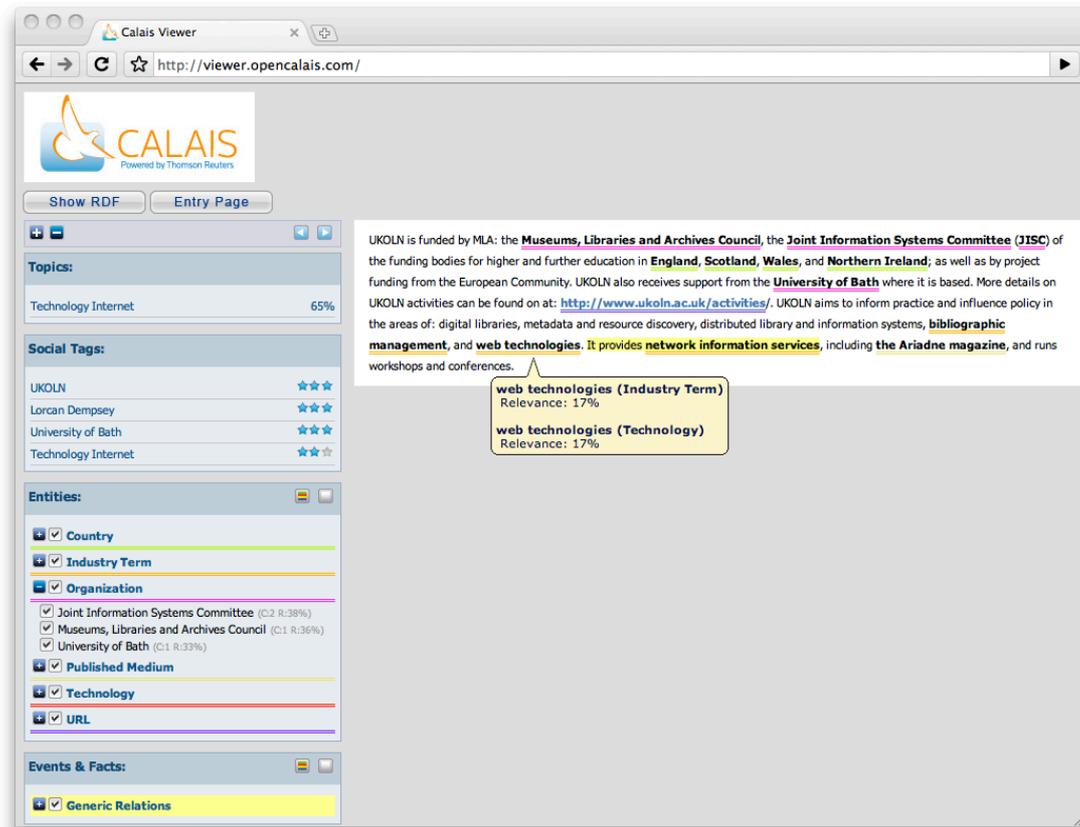
In a simple illustration, I copied the first paragraph of UKOLN's 'About' page,<sup>61</sup> stripped out the URLs, and pasted it into the Calais Viewer tool to achieve the result below.

---

<sup>59</sup> <http://opencalais.com/>

<sup>60</sup> <http://d.opencalais.com/er/company/ralg-tr1r/9e3f6c34-aa6b-3a3b-b221-a07aa7933633.html>

<sup>61</sup> <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.

### Freebase

San Francisco-based Freebase<sup>62</sup> is a community-maintained 'free database of the world's information,' backed by significant venture capital.<sup>63</sup> 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<sup>64</sup>.

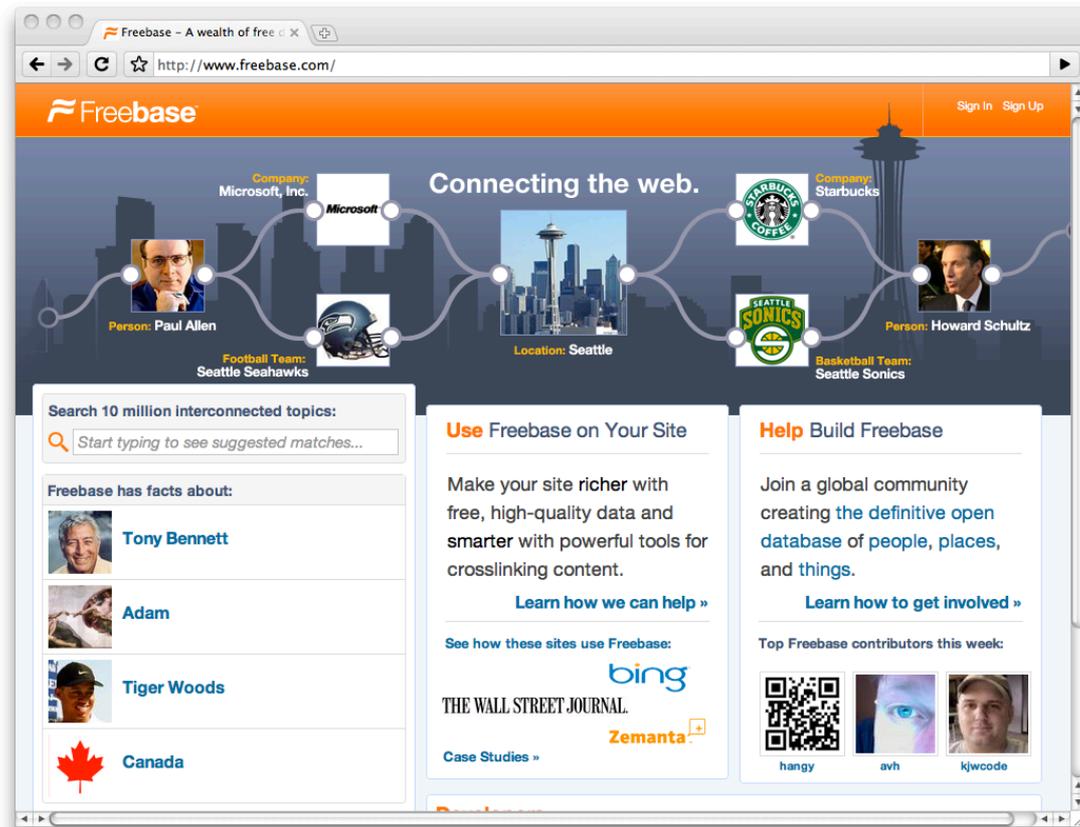
Towards the end of 2008 Freebase launched<sup>65</sup> a new RDF service<sup>66</sup> that enabled responses to api calls to be returned in RDF, making Freebase content available to those building Linked Data applications.

<sup>62</sup> <http://www.freebase.com/>

<sup>63</sup> <http://web2innovations.com/money/2008/01/18/massive-second-round-of-funding-for-freebase-42-million/>

<sup>64</sup> [http://en.wikipedia.org/wiki/Semantic\\_Web\\_Stack](http://en.wikipedia.org/wiki/Semantic_Web_Stack)

<sup>65</sup> [http://blog.freebase.com/2008/10/30/introducing\\_the\\_rdf\\_service/](http://blog.freebase.com/2008/10/30/introducing_the_rdf_service/)



Freebase

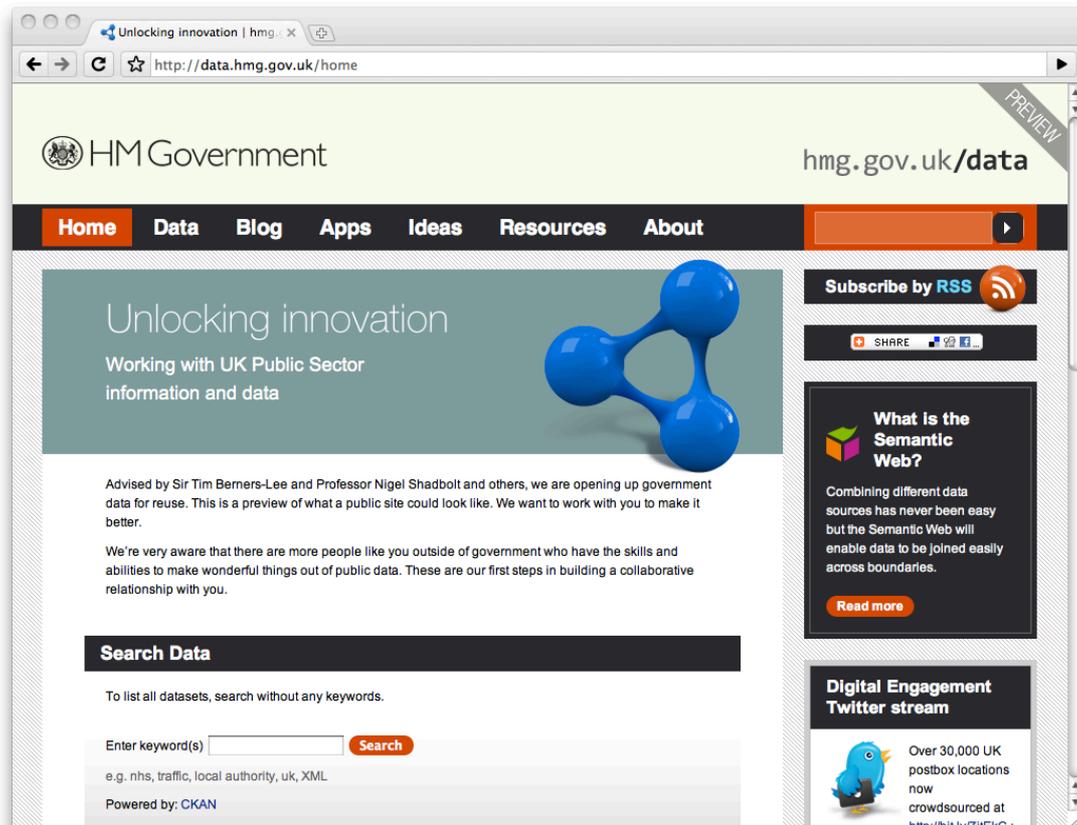
### UK Government

Prime Minister Gordon Brown announced<sup>67</sup> 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.

A number of data sets are already available on the Government's beta site, [data.hmg.gov.uk](http://data.hmg.gov.uk), and this month's *Putting the frontline first* document reiterates the promise that plenty more will follow. As well as simplifying access to previously 'available' data, lobbying behind the scenes appears to have been successful in changing attitudes to data from Ordnance Survey, the Post Office and other agencies that previously charged significant fees for access to their data.

<sup>66</sup> <http://rdf.freebase.com/>

<sup>67</sup> <http://www.guardian.co.uk/technology/2009/jun/10/berners-lee-downing-street-web-open>



The UK Government data site, currently under development

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 and compelling case studies emerge to illustrate the real value 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*<sup>68</sup>. 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 are more promising than falling back on the traditional methods of procuring yet-more proprietary silos.

Although not directly mentioned in *Putting the frontline first*<sup>69</sup>, 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 institutions 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<sup>70</sup>

"...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)

<sup>68</sup> <http://www.bis.gov.uk/policies/higher-ambitions>

<sup>69</sup> <http://www.hmg.gov.uk/frontlinefirst.aspx>

<sup>70</sup> <http://research.microsoft.com/en-us/projects/orechem/>

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

Elsewhere, Oxford's BRIL<sup>71</sup>

“...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.”

(BRIL project description)

Ben O'Steen 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. RDFa markup in the pages makes the content available for further processing in third party applications where required.

---

<sup>71</sup> <http://brii.ouls.ox.ac.uk/>

# Recommendations for Future Work

There are ten recommendations for future work to exploit the opportunities offered by Linked Data. These fall into three broad areas, and are detailed below.

## 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, and providing lightweight hooks to begin combining data from otherwise incompatible systems.

**Recommendation 1:** review Cabinet Office guidance<sup>72</sup> on the creation of URIs for the UK public sector and W3C guidelines on 'Cool URIs<sup>73</sup>.' 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 a new 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<sup>74</sup> Data Incubator<sup>75</sup> and vocab.org<sup>76</sup> provide infrastructure that could be used in converting these core resources to more

<sup>72</sup> [http://www.cabinetoffice.gov.uk/media/308995/public\\_sector\\_uri.pdf](http://www.cabinetoffice.gov.uk/media/308995/public_sector_uri.pdf)

<sup>73</sup> <http://www.w3.org/TR/cooluris/>

<sup>74</sup> <http://blogs.talis.com/nodalities/2009/05/growing-the-web-of-data-with-data-incubator.php>

<sup>75</sup> <http://dataincubator.org/>

<sup>76</sup> <http://vocab.org/>

accessible forms. An earlier version of the JACS codes used to describe undergraduate courses is available via DataIncubator<sup>77</sup>, for example, and might usefully be built upon.

**Recommendation 2:** identify the core set of widely used identifiers (JACS codes, institutional identifiers, etc) and facilitate 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<sup>78</sup> 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 BRIL. Web specifications such as FOAF<sup>79</sup> 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<sup>80</sup> 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.

**Recommendation 3:** explore the value of providing infrastructure that members of the community may use in hosting a uniquely identified personal profile, linked to institutional, professional and social network identities as appropriate. Assess whether or not the existing Names infrastructure offers a viable foundation upon which to build.

---

<sup>77</sup> <http://jacs.dataincubator.org/>

<sup>78</sup> <http://names.mimas.ac.uk/>

<sup>79</sup> <http://www.foaf-project.org/>

<sup>80</sup> <http://foafbuilder.qdos.com/>

## 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.

**Recommendation 4:** evaluate the effectiveness of the Office of Public Sector Information (OPSI) Unlocking Service<sup>81</sup>, and consider whether a similar approach might be of value in helping the community identify data sets to prioritise. Take the lead in lobbying for access to frequently requested resources, and consider the use of financial incentives to facilitate access where required.

Once identified, there is clearly 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.

**Recommendation 5:** evaluate the effectiveness of existing community efforts such as Data Incubator<sup>82</sup>, and establish a register of individuals and organisations able to help convert data. Allocate funding to a finite set of data conversions, prioritising proposals that demonstrate a combination 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, 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.

---

<sup>81</sup> <http://www.opsi.gov.uk/unlocking-service/OPSI/page.aspx?page=UnlockIndex>

<sup>82</sup> <http://dataincubator.org/>

**Recommendation 6:** fund a focussed piece of work to validate existing data licenses such as those from the Open Data Commons. Actively engage with Government work on data licensing. Use existing JISC Advisory Services to disseminate findings, and evaluate the feasibility of high level endorsement for an Open Data approach.

The emphasis of Linked Data activities tends to be placed upon large data sets, but there is equal value in enriching web pages by embedding structure with RDFa<sup>83</sup> -- especially now that Yahoo! SearchMonkey<sup>84</sup> and Google Rich Snippets<sup>85</sup> explicitly crawl. RDFa is particularly useful in adding explicit structure to descriptive content, and might usefully be used in everything from OPAC pages describing a book to the University vacancies site. Mark Birbeck, who originally proposed RDFa, recently completed work to add RDFa to job information provided by all central Government departments<sup>86</sup>. 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.

**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<sup>87</sup> in supporting exposure of course data to the web.

---

<sup>83</sup> <http://en.wikipedia.org/wiki/RDFa>

<sup>84</sup> [http://developer.yahoo.net/blog/archives/2008/09/searchmonkey\\_support\\_for\\_rdfa\\_enabled.html](http://developer.yahoo.net/blog/archives/2008/09/searchmonkey_support_for_rdfa_enabled.html)

<sup>85</sup> <http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=99170>

<sup>86</sup> <http://webbackplane.com/mark-birbeck/blog/2009/04/23/more-rdfa-goodness-from-uk-government-web-sites>

<sup>87</sup> <http://www.xcri.org/>

## Supporting Measures

Independently to this report and the associated Funding Call expected early in 2010, the JISC-funded CETIS has been addressing semantic technologies and – increasingly – Linked Data through its Semantic Technologies Working Group<sup>88</sup>. At a meeting in December 2009, this group recommended extending its charter with a programme of work for 2009.

The scope of these two activities is not exactly the same, and there is no significant issue in allowing both to continue.

It will be important to ensure that there is active communication and cross-fertilisation in order to maximise exploitation of any synergies that emerge, and to avoid unnecessary duplication of effort.

**Recommendation 8:** explicitly establish internal communications to ensure that activities funded by JISC’s Infrastructure & Resources Committee (JIR) are aware of and do not unnecessarily duplicate those started under the auspices of the JISC Learning & Teaching Committee (JLT) and CETIS.

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 9:** evaluate the models developed for existing web data services. Consider ways in which the community can *use* existing data infrastructure rather than reinventing it, and *enrich* existing data infrastructure rather than competing with it.

---

<sup>88</sup> [http://wiki.cetis.ac.uk/Educational\\_Content\\_semtec](http://wiki.cetis.ac.uk/Educational_Content_semtec)

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 10:** identify a focus for Linked Data activities, perhaps within an existing JISC Advisory Service. Consider funding attention-raising activities such as a competition for compelling uses of Linked Data within the community.

DRAFT

# Acknowledgements

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

Many thanks to everyone who willingly contributed time to share their perspectives.

Mark Birbeck, webBackplane

Rachel Bruce, JISC

Ken Chad, Ken Chad Consulting

Chris Clarke, Talis

Keith Cole, Mimas

Adam Cooper, CETIS

Leigh Dodds, Talis

Jim Downing, University of Cambridge

David Flanders, JISC

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

Thanassis Tiropanis, University of Southampton

Paul Walk, UKOLN

Jo Walsh, EDINA

I would also like to thank attendees at the SemHE workshop<sup>89</sup> in Nice and participants involved in CETIS' December meeting of their Semantic Technologies Working Group<sup>90</sup>.

---

<sup>89</sup> <http://www.semhe.org/>

<sup>90</sup> <http://jisc.cetis.ac.uk/events/register.php?id=211>