Leeds Linked Data Business Case

Andy Turner

Metadata

- Draft version 0.8 (2011-07-26)
- Stable 0.8 version available as a PDF via:
    - http://www.geog.leeds.ac.uk/people/a.turner/personal/blog/archive/2011/07/26/LeedsLinkedDataBusinessCase.pdf
- This document was created using Google docs
    - If you are reading an exported version, then consider that there may be a more recent version on the world wide web available via the following URL:
        - https://docs.google.com/document/d/1iUMWsoPeP3DnYvj0tQILB7DzqHiduEE8JD6vwsS_Hj4/edit?hl=en_US
- If you edit the document, please include your name in the author list under the title and increment the version number.
- This document is intended to raise awareness and support for developing the University of Leeds (UoL) linked data and is aimed specifically as a supporting document for the Vice-Chancellor's Executive Group (VCEG) and the Strategy Group
    - This is because support from top is important. It can be used to overcome social and organisational barriers that may otherwise delay further the development of linked data. Additionally, top level management can decide to expend more resource on the development of linked data and the enabling e-Infrastructure.

Contents

1. Introduction

This document aims to detail a business case for the development of linked data about Leeds and the University of Leeds (UoL) without going into technical details. A brief overview Case for Support document has also been drafted which skips much of the detail provided herein [1]. That case support focusses on the university itself although it introduces as does this the broader context in which the university operates. This document also does not go into technical implementation detail, but outlines why support is needed and suggests what form might take and how implementation may proceed.

It can be argued that much of the UoL linked data should be Linked Open Data (LOD). Making data open, as open as is legally possible and ethically reasonable to make it, is key to maximising the utility of the data. The best uses for data are often not thought of by the custodians of the data themselves. Changes to the way students are registered and graduate and the way staff are hired and leave posts will be needed to ensure that the university can keep track of them to help measure the impacts they have and the university has, and to provide information to prospective students and people seeking employment. This is in-line with: the recently released BIS report “Higher Education: Students at the Heart of the System” [2]; the RCUK Excellence With Impact Web Site which focuses on Knowledge Exchange an Impact (KEI) [3]; and, the joint HEFCE, RCUK and GuildHE reports “Provision of information about higher education: Outcomes of consultation and next steps” which focuses also on Key Information for Students (KIS) [4].

Some of the potential benefits of linked data and LOD are suggested in this business case, and issues to do with implementation and risks are introduced. This document was drawn up following a meeting with the University of Leeds Web Team on 2011-06-23 shortly after which a placeholder for the open public facing linked data for the UoL, *Leeds LOD*, was made available on-line via http://data.leeds.ac.uk [5].

The business case is presented as clear cut, which is both bold, but also if this comes across, then it will be successful. The aim of this document is to convince the UoL management that there is no question of if this needs to be done, more one of how do we do it and what are the next steps. The UoL is developing its data and it is becoming more linked and integrated over time, but there is a step change between letting this happen organically and catalysing it with support to evolve a new integrated data system. Much of the data will be wanted as publicly available - free at the point of use - data or Linked Open Data (LOD). Implementation and organisational issues to do with the development of linked data go hand in hand. With support from the top and through all levels of the organisation and with the key stakeholders in the Faculties of Engineering and Environment, Leeds University Business School (LUBS), the Library and Information Systems Services (ISS) on board how can we fail? If the Open University, the University of Southampton, University College London, the University of Oxford, the University of Edinburgh, the University of Bristol and the University of Lincoln are convinced and making strides towards this, then they must have convincing business cases supported from the top also. We are not pioneering this at the bleeding edge which gives us the advantage of learning from others and following a straighter implementation path, but it also means that we do not get research kudos in this field and may ultimately miss out on some of the funding for this from the Joint Information Systems Committee (JISC).

Arguably the biggest risk to the UoL is to not fully appreciate the importance of linked data and LOD and in being slow to adopt, our impact measures will lag making it hard for us to get the credit we deserve. One could argue that, a lack of resourcing for this now, may be costly in the long run, but we have impetus for this and good will. The battle for hearts and minds is being well won and there is growing support for this proposal across the boards. Is the UoL satisfied with survival in the top half, or would it rather be seen to be amongst the top 10 in 10 years time? It can be argued that resources put into this in the right way will draw commensurate rewards in the medium term. We have the will, we know the way, we are making progress, but we could make it faster if we were better resourced for this specifically. Additionally, people get more expensive over time and although better tooling will come to our aid in developing our data, getting the data organised will be a large and time consuming and reasonably manual challenge, so the sooner we start, the better.

There is little need for capital investment to support the development of linked data and LOD. What is likely to be required, more than anything else is a team that focuses on this and which includes developer, implementers, academics, library, manager and other support persons. The idea is to draw all the data into the new database. Things do not need to change much, there is no manifesto for change associated with a move to support linked data, we just need to hold on to our values and be careful not to focus on short term revenue and cost cutting at the expense of longer term revenue and enhancement.

Trailing in the wake of too many universities will not help our World Class ambition. Like we needed web pages to be developed in the 1990s, we now need linked data.

Our web content is generally of a high standard, but it is geared for navigation by human users albeit with the help of search engines. We now need a specially organised collection of web content that allows for specialised searches that pull the data and allow it to be aggregated, formatted into triple stores [6], analysed, generalised and visualised in many different ways. Providing access to this data to external organisation is key for establishing collaboration and driving impact and is being increasingly mandated by government [3].

A briefing document prepared for the Web Team meeting on 2011-06-23 provides a reasonably detailed introduction to linked data and LOD [6]. Perhaps a better introduction is available as the final formal blog post of the Lucero project [7]. More recently Christopher Gutteridge from the University of Southampton School of Electronics and Computer Science Web Team addressed the difference between linked data, open data and Resource Description Framework (RDF) Data [8]. A definition of linked data, as provided by the UK government; and, a definition of open data, from Wikipedia - are provided below:

> “Linked data is data in which real-world things are given addresses on the web (URIs), and data is published about them in machine-readable formats at those locations. Other datasets can then point to those things using their URIs, which means that people using the data can find out more about something without that information being copied into the original dataset.” [9][1]

> “Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other

[1] The Uniform Resource Identifier (URI) link has been added to this quotation.

> mechanisms of control. While not identical, open data has a similar ethos to those of other "Open" movements such as open source, open content, and open access. The philosophy behind open data has been long established (for example in the Mertonian tradition of science), but the term "open data" itself is recent, gaining popularity with the rise of the Internet and World Wide Web and, especially, with the launch of open-data government initiatives such Data.gov." [10]

The composite of linked data and open data is Linked Open Data (LOD) which for the greater part is what the UK government are driving for in terms of enhancing our democracy through transparency in decision making [11]. Not all data though is immediately open data, it can be embargoed and restricted for security reasons, but ultimately data is likely to be released into the long term to the public domain as with most UK government records.

Section 2 outlines the potential benefits of linked data and linked open data. Section 3 outlines potential implementation risks and mitigation strategies for preventing risks and dealing with problems. Section 4 introduces requirements and reflects on where we are and what are the next steps.

2. Potential benefits of Linked Data and LOD

Perhaps the biggest benefit is that well developed LOD will allow others to identify and measure the impacts of our education and research. It should allow expertise to be identified and make it easier for prospective students to identify the right courses for them.

As research becomes more computational and data intensive then software, models and data will become as important as peer reviewed journal publications. Also the way in which scholarly communication is taking place is changing from a more paper based and slow peer reviewed system, to a more rapid information and communication technology driven blog style communication which promotes conversation with the public and other stakeholders that are interested in the research. Publication as a form of communication is changing. Reviewers are more able to review in the open and post and pre the perhaps more official peer review system that it will ultimately replace. This means that reviewers reviews can also be credited as more aspects of the publication are laid open for metrics to be generated

The ClimateGate scandal at UEA in 2009 has lead to researchers and the public demanding access to both publicly funded research data and the algorithms used to transform that data into information and knowledge [12]. Funding councils, such as EPSRC, are demanding that publications derived directly from funded research must be made available as openly accessible documents [13] [14]: it is likely that they will soon demand similar access to data, algorithms and other outputs from the research process.

Keeping track of what we help to develop and what services we provide and how these are used will be key for alternative impact metrics. The university will look after some data in institutional repositories, but much data will be archived and accessed from elsewhere and platform further research. LOD about the provision of capability and the use of resources for research and the research data will be increasingly important. We need to nurture and develop our e-Infrastructure for research, and linked data is an important part of this.

For any large organisation operating in the digital economy, linked data is growing in importance. Linked data for universities includes data about people (staff and their roles, skills, experience, qualifications; and non-staff), hard and soft infrastructure (facilities and sub-organisations), research and courses. The linked data should link outwith the university to detail relationships with other organisations and individuals and link to informative and trusted authoritative endpoints.

For strategic and operational management, linked data can help inform and act as a tool for understanding and developing the function and form of the organisation. The importance of people coming up with good ideas for evolving the organisation and putting plans into action should not be down played, but similarly, the potential of linked data should be appreciated then realised. The University of Leeds is part of the information society, let us make it more so and stand out as a world class institution for developing knowledge and having positive impacts on society and the environment.

A key to understanding and developing a business - is mapping and appreciating and driving efficiencies in its use of data.

A big issue in opening up linked data is to identify how open different data should be. What data should be made open and how? This may all depend on the use of the data and licensing is critically important as are roles and groups and appropriate access controls. These issues are elaborated on in Section 3. In terms of maximising benefit, the more open the data is made, the more uses it can have, and the more benefits there are that can be realised from it. Much data will have little or no confidentiality or copyright issues that prevent it being released to the general public under an appropriate non-attributing open data license. Much data will have intrinsic commercial value. However, the university should resist the notion of selling these data for profit, or imposing too many restrictions on its license and usage conditions. In general, the more open the data is the more likely it is that others will use it and add value to it. If we are indeed keen to measure impact, then linked data and LOD are likely to be at the heart of doing this well.

In the UK government Higher Education (HE) reform White Paper [2] numerous metrics are called for and the Higher Education Funding Council for England (HEFCE) has jumped to address this, not least with respect to the Key Information Sets (KISs) [16], [17]. The KISs are to allow prospective students to get standard information about the organisation, costs and expected benefits of degree courses from different UK HE providers. The UKRC Report of the e-Infrastructure Advisory Group 2011 also calls for open information about collaboration between universities and other organisations [18]. Arguably the best way to deliver this and keep track of collaboration is via a united effort in developing data.ac.uk of which data.leeds.ac.uk should be a very important part.

LOD is a fairly nebula concept and it reaches beyond academia into government, quasi and non-government organisations, and commercial business enterprises. Links with local government, public service organisations, libraries museums and archives (LMA) and other educational institutions is key. Much of this regional work is geographical in nature and the development of a shared geospatial information system should begin. To this end, the development of Leeds LOD is being done geographically feeding our own teaching and research and through the provision of a new undergraduate geography module “Leeds: From Local to Global” convened by David Bell. Quite how we manage to feed this activity into teaching and research will depend as much on the willingness of academic staff to engage in the writing and provision of new course material and for the staff and students to engage with others in the local community.

In summary, the key benefits to realise are in developing detailed information about impact. The impact of our research, education and staff development reaches far and wide through our alumni, our scientific and non-scientific output, through our former staff that have moved on to vital roles employed in other organisations and through our impacts on the places, organisations and people we interact with both formally (through collaboration agreements, secondments and visits) and informally (through all other forms of collaboration not least being knowledge development). We not only need to be bringing in funding and turning this into products, but we also need to be demonstrating value for money. Linked data should enable us to perform a more detailed cost-benefit analysis for research funding that goes beyond dividing output by income and measuring the general weights of the numerator and denominator. Linked data will enable us to improve operational efficiency and make it easier for users to use, developers to develop, and maintainers to maintain our facilities, organisational information and educational and research content and other resources. We should not forget the importance of us having our own house in order as we try to get more involved in developing a Geographical Information System and LOD for the region.

3. Potential implementation risks and mitigation strategy

Access control is vital so that data is only accessed by those authorized to access it and made available to consumers that can make good use of it. A role based access control is probably part of this, but access control for some data needs to be fine grained enough to work at the individual and specific purpose level. We are of course bound by data protection legislation and to some extent, the details about individuals, particularly those that are not staff should be to some extent opt in although encouraged. Minimising disclosure risks by not opening up data is itself a risk. By default data should be made more open access unless there is a legal reason not to. The data will increase in value through use, but we should resist the temptation to put intellectual property constraints and restrictive licensing in place and to further restrict use by attempting to sell data wherever possible.

Metadata, data about the data will be key. It is expensive to develop and chould be developed to adhere to the appropriate standards. Metadata is integral to data security, but like security it should be viewed as enabling. In many cases, the metadata will outweigh the data. It can grow to include details of the use of the data, but this can only easily be done with appropriate licensing and by having in place an accounting mechanism that records how the data are accessed. Licensing and monitoring are key, perhaps the former is most important and again getting this right is enabling and helps us understand what is being done with the data. Notwithstanding that our LOD will be sucked up into clouds and accessed from these copies making it difficult to keep a track of all that it is used for (presuming we would want to).

Effects on existing services and business practice are inevitable. Available LOD is copied from (pulled or sucked) out for use as well as being crawled for indexing and to identify changes. If a lot of this activity goes on simultaneously, there are load balancing issues. Being able to restrict the pull of data is important, especially if other services are dependent on the same computational infrastructure providing the LOD data to the world. Watching and keeping a track of the use of the data, where it is going and how it is searched by crawlers is important for helping us learn what happens to it. Failing to do this from the very start is something we may learn to regret. Again appropriate metadata is important, particularly versions for data and metadata, and dates of last and planned modifications.

It is people that mainly develop our data and there is a risk of upsetting them if implementation is too heavy handed. Some people will be using a system that despite flaws, essentially works, they will be familiar with it and some would find it difficult to learn and look after the data another way. Others will be keen to gradually move from working with the tools and work-flow they know, to a new system. The key thing is to map where the data is and how it is updated so that we can pull all the data and link it in the RDF format [19] [20]. Gradually, we can work towards a fully integrated system. What we do not want is to create another copy or version of data necessarily, key is integration and getting all parts using and developing the same data. Also this should be application lead, we do not necessarily want to be linking data for the sake of it, we should have implementation use cases in mind, such as broadcasting news about a sub-organisation or individual activities. We need to map the canonical data and work with the data developers and custodians without them fearing losing control of the parts of our data they look after and may consider as their data, not to be shared. Tact and diplomacy will be key and dealing with clashes of personality could be important.

There are resourcing issues to do with developing linked data. We need skills and developer and implementer time. More importantly, we need to work together and finding the time to work together can be difficult as people are stretched already and find it difficult to commit to more work and more meetings in the short term even if this would save time in the longer term. Under resourcing this is not really an issue, the main barrier is not likely to be technical or because we do not find time to do it. It is more likely to be political and organisational. This is why support from the top is so important. The risks of alienating staff and creating division necessarily have to be avoided and being open and collaborative and having support at all levels of the university is key to mitigating this.

4. Requirements

Documentation for implementation and a road map for developing Leeds Linked Data would be good things to have to reduce reliance on key individuals and to make implementation less risky and more sustainable in the long term. Requirements gathering could be the next step, but so long as we set off in a scalable way, then getting going with producing LOD and monitoring its use is perhaps key. LOD should be developed incrementally and we should use and develop open source tooling for this. There are few requirements for us to get going. We are not exactly waiting for a green light. The next steps are to develop the relationships and a team that reaches across the Faculties of Environment and Engineering, the library and Information Systems and Services (ISS). One project we should bear in mind with all this is DART which has been addressing the issues that are likely to surface with this in a multi-institutional context [21].

5. References

1. Leeds Linked Data: Case for Support Google docs Document https://docs.google.com/document/d/1QdQXRJUQqW2Ug7ewihYAv1Ij4W3JbrT2dMrp9AANYq4/edit?hl=en_US&pli=1
2. BIS (2011-06) Higher Education: Students at the Heart of the System. Presented to Parliament by the Secretary of State for Business, Innovation and Skills http://discuss.bis.gov.uk/hereform/white-paper/
3. RCUK Excellence With Impact Web Site http://www.rcuk.ac.uk/kei/
4. HEFCE (2011-06) Provision of information about higher education: Outcomes of consultation and next steps. Joint report from HEFCE, Universities UK and GuildHE Provision of information about higher education Outcomes of consultation and next steps http://www.hefce.ac.uk/pubs/hefce/2011/11_18/11_18_35454121.pdf
    - The document sets out how HE organisations “intend to improve the accessibility and usefulness of information about higher education courses, from September 2012. It also sets out how the National Student Survey will be developed, and what wider information should be made available by universities and colleges.”
5. The University of Leeds Linked Open Data Web Site http://data.leeds.ac.uk
6. Wikipedia Triplestore Article http://en.wikipedia.org/wiki/Triplestore
7. Developing and using data.leeds.ac.uk as (open) (geospatial) linked data: A briefing document for the University of Leeds for a Web Team meeting on 2011-06-23 https://docs.google.com/document/d/1EgzYQKK1iq5SMS1Hn1hOfVO7yvJrGYESpxZCBPIOmwc/edit?hl=en_US
8. The LUCERO project: Linking University Content for Education and Research Online: What is Linked Data? Web Page http://lucero-project.info/lb/what-is-linked-data/
9. University of Southampton, School of Electronics and Computer Science Web Team, Christopher Gutteridge, Linked Data vs Open Data vs RDF Data Blog Post http://blogs.ecs.soton.ac.uk/webteam/2011/07/17/linked-data-vs-open-data-vs-rdf-data/
10. Wikipedia Linked Data Article http://en.wikipedia.org/wiki/Linked_data
11. Wikipedia Open Data Article http://en.wikipedia.org/wiki/Open_data
12. HM Government: data.gov.uk Beta: Opening up Government: Linked Data Web Page http://data.gov.uk/linked-data
13. BBC (2009-12) 'Show Your Working': What 'ClimateGate' means http://news.bbc.co.uk/

1/hi/8388485.stm

14. EPSRC (2011) Policy on access to research outputs http://www.epsrc.ac.uk/about/infoaccess/Pages/roaccess.aspx
15. RCUK (2006-06) Research Councils UK' updated position statement on access to research outputs http://www.rcuk.ac.uk/documents/documents/2006statement.pdf
16. HEFCE Key Information Sets Web Page http://www.hefce.ac.uk/learning/infohe/kis.htm
17. HEFCE Providing Information about Higher Education Web Page http://www.hefce.ac.uk/learning/infohe/kis.htm
18. UKRC Report of the e-Infrastructure Advisory Group 2011 http://www.rcuk.ac.uk/research/xrcprogrammes/eInfrastructure/Pages/home.aspx
19. Wikipedia Resource Description Framework Article http://en.wikipedia.org/wiki/Resource_Description_Framework
20. World Wide Web Consortium: Semantic Web: Resource Description Framework Web Page http://www.w3.org/RDF/
21. The DART Project Web Page http://dartproject.info