Leeds Linked Data Business Case

Andy Turner

- Draft version 0.7 (2011-07-08)
- In draft form this document is not open...
- 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 at the Vice-Chancellor's Executive Group (VCEG)
    - 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 present a business case for the development of linked data about the University of Leeds (UoL) without going into the technical details. Much of the UoL linked data should be Linked Open Data (LOD). Making data as open as is legally possible to do so is key to maximising their utility. 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 it is having and to provide information to prospective students and people seeking employment. 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 [1] 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 [2].

The business case is clear cut. The UoL has no option but to develop linked data and LOD. It is perhaps more of a matter of how it will organise itself to do so. Arguably the biggest risk is to not

fully appreciate the importance of linked data and LOD and then to not invest commensurately and gain only small amounts of benefit. Failure to resource this now will probably be more costly in the long run. People get more expensive over time and although better tooling will come to their aid, getting the data organised will be a large and time consuming challenge.

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 of developers that work with others on the data.

Trailing in the wake of too many universities will not help our World Class ambition. There is no option, like we needed web pages in the last decade that could be navigated 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 analysed, generalised and visualised in various ways.

For a reasonably detailed introduction to linked data and LOD, please refer to the briefing document prepared for the Web Team meeting on 2011-06-23 [3]. Perhaps a better introduction is available as the final formal blog post of the Lucero project [4]. 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.” [5][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 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.” [6]

The composite of linked data and open data is Linked Open Data (LOD).

---

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

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, software, models and data will become as important as peer reviewed journal publications. 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 [7].

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 references.

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 [4] numerous metrics (Key Information for Students [8]) are called for [9]. These are to allow prospective students to get standard information about the organisation, costs and expected benefits of degree courses from different UK HE providers. It also calls for open information about collaboration between universities and other organisations [10]. 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.

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 money 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 and general investment in our higher education institute. 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 develop GIS 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.

Metadata, data about the data will be key. It is expensive to develop and needs 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 or by watching the data access and generalising this information. 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 and throttling the pull of data is important especially if other services are depended upon on the computational infrastructure. Watching and keeping a track of the use of the data, where it is going and how it is searched by crawlers is important as we learn what happens. 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 rubbing them up the wrong way if implementation is too heavy handed. Some people will be using a system that despite it's flaws, works, they are familiar with it and they 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 workflow 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 suck and link the data internally and format it as RDF/XML. Gradually, we can work towards a fully integrated system. What we don't want is to create another copy or version of data necessarily, key is integration and getting all parts using and developing the same data. We need to map the canonical data and work with the data developers and custodians without treading on their toes. 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 time and finance. The more resource we gear into this the better. The risk of under resourcing this is that corners might be cut and things get a little messy, however, no matter how little resource there is, the main barrier is not likely to be technical, it will be political and this is why support from the top is so important. As a last resort, developers will have to pull rank and request support. The risks of alienating staff and creating division necessarily have to be avoided and being open and collaborative 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 the operation 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.

5. References

1. https://docs.google.com/document/d/11HHYRmknrUr65nAdFO9WBEmFVdPWc3FbUvCpQzCyqSg/edit?hl=en_US&authkey=CPjDl-wN
2. https://docs.google.com/document/d/11HHYRmknrUr65nAdFO9WBEmFVdPWc3FbUvCpQzCyqSg/edit?hl=en_US&authkey=CPjDl-wN
3. http://lucero-project.info/lb/what-is-linked-data/
4. http://www.bis.gov.uk/HEreform (http://discuss.bis.gov.uk/hereform/white-paper/
5. http://data.gov.uk/linked-data
6. http://en.wikipedia.org/wiki/Open_data
7. http://www.rcuk.ac.uk/research/xrcprogrammes/eInfrastructure/Pages/home.aspx
8. http://www.universitiesuk.ac.uk/Newsroom/Media-Releases/Pages/keyinformationforstudents.aspx
9. http://www.universitiesuk.ac.uk/Newsroom/Media-Releases/Pages/Responsetogovernmentwhitepaperonhighereducation.aspx
10. http://www.rcuk.ac.uk/media/news/2011news/Pages/110628.aspx