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

(Version 1.0 2011-06-21)

Andy Turner

Contents

1. Introduction

1.1. What is Linked Data?

“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.” [1][1]

“Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF."”[2][2]

A introductory guide to developing linked data developed by colleagues at the University of Southampton is available online [3]. Much Web content is linked data. Structured data about a university made available on Web pages is key to developing its Linked Open Data (LOD). The more links it has the better, and the more these links lead to canonical endpoints (URIs), the better. Design issues for Linked Data are worked through by the World Wide Web consortium W3C in collaboration with other standards defining organisations [4].

1.2. What is Open Data?

“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.” [5]

There is a manual on creating open data [6], which addresses the definition of open data [7] and which refers to the open definition developed by the open knowledge foundation [8].

One of the key things about Linked Open Data (LOD) is the use of what are known as cool URIs - URIs that don’t change [9] and WebIDs (coolURI end points) for unambiguously referring to real world things, such as organisations and individual people [10].

1.3. What is Geospatial Data?

Geospatial data is spatially referenced to a geoid location, and often also temporally referenced.

---

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

[2] Further links have been added to this quotation.

In the context of this document, we are interested in data about planet Earth including observations and measurements. Data collected about people, organisations and places is often geospatial and directly or indirectly linkable with other geospatial data. Direct linkage tends to use spatial geometry and may also use temporal date based referencing. Indirect referencing may use an an attribute for a physical address or geographical identifier of some kind. This could be the name of a place or a physical event identifier (physical events are inherently spatial and temporal and their specification in these terms allows data about them to be linked geospatially). Linkage is commonly done using; place names (e.g. Leeds, Swinsty Reservoir, The School of Geography East Building), postal and telecommunication addresses (or parts thereof e.g. a postcode, or a telephone area code). A social network hashtag for a physically located conference or the URL for an individual post which is geolocated could be used as an indirect geographical reference.

Spatial reference can take various forms, commonly coordinates are used to represent points, lines and polygons in a geometrical system. The geometry is spatially two or three dimensional and positioned and orientated relative to some well defined geoid with origin and axes of coordination. For Earth, there are commonly used definitions for its centre through which there is an axis based on rotation and an orthogonal plane for the equator. A point on the surface of the Earth is also defined for longitude and temporal reference. Often a projection is used to allow for representing coordinates locations in two dimensions on Earth's surface. The geometrical precision and accuracy and the details of temporal referencing can vary considerably.

Geohashs can be used to collapse the dimensions of reference for a point in two dimensions using data interleaving [11]. This concept can be extended to collapse spatially three dimensional data and also spatial data incorporating time into string representations that's similarities can provide information on proximity.

Part of the special nature of geospatial data is that it is inherently linkable through coincidence.

1.4. Progress elsewhere

1.4.1. LOD for other universities

It is to be expected that most universities are forging ahead developing their LOD. Currently there are several sharing their experience and developing in a collaborative fashion of sorts and which are listed on data.ac.uk (http://data-ac-uk.ecs.soton.ac.uk/) [12]. The first steps in developing LOD for universities have been taken and the path to take is becoming clear. However, new developments are continually occurring and this promotes an agile approach.

Pioneering work has been done in an open and collaborative manner by inspirational people based at the Open University and the University of Southampton. Various projects are involved as are other academic organisations and other organisations providing services and software and know how. It would not surprise me if the University of Oxford goes live very soon with something we can all aspire to. At the moment their data seems to be held privately.

A list of data.ac.uk (http://data-ac-uk.ecs.soton.ac.uk/) URLs to watch (some of which are not live links yet):

- data.open.ac.uk [13]
- data.soton.ac.uk [14]
- data.online.lincoln.ac.uk [15]
- data.ox.ac.uk [16]
- data.kcl.ac.uk/ [17]
- data.ed.ac.uk [18]

1.4.2. LOD for other related organisations

Governments, particularly the UK [19] and USA [20] governments are opening up their data about their organisations and about what they do as LOD. In June 2011, the UK government released an organogram [21] which details the structure of its internal working. A generic effort to map all government organisations is also available on-line [22]

Another group developing LOD are libraries, museums and archives. CLAROS is working to enable simultaneous searching of major collections in university research institutes and museums [23]. The JISC funded LOCAH Project [24] involves Mimas and UKOLN working together in partnership with Eduserv, Talis and OCLC to make Archives Hub [25] and Copac [26] data available as structured Linked Data, for the benefit of education and research. The project aims to "place archival and bibliographic data at the heart of the Linked Data Web, enabling new links to be made between diverse content sources and enabling the free and flexible exploration of data so that researchers can make new connections between subjects, people, organisations and places to reveal more about our history and society" [24]. "The Copac® library catalogue gives free access to the merged online catalogues of many major University, Specialist, and National Libraries in the UK and Ireland, including the British Library" [26]. University of Leeds Library is a contributing library [27]. Europeana is a large collaborative project co-funded by the European Union and involving upwards of 100 partners to develop LOD for paintings, music, films and books from European galleries, libraries, archives and museums [28,29].

1.4.3 Geospatial LOD

There is at least one effort to link together OpenStreetMap, Geonames and DBPedia data as geospatial LOD or linkedgeodata [30]. There is a vast amount of open geospatial data (a geocyberspace too vast to concisely summarise and link to here) which given its spatially and temporally referenced nature is inherently linkable.

The UK National Mapping Agency, the Ordnance Survey, also develop some Linked Open Data which offers a means for disambiguating references and relating information about physical entities represented in geospatial data about the UK [31,32].

2. Data about the University of Leeds

Our organisation (Leeds) is an academic higher education institution (geared for teaching and

research). Based on physical infrastructure it provides a place of work and study to many sub-organisations and individual people [33]. Leeds has a main campus on the edge of Leeds city centre. The university and many of its sub-organisations link with other organisations and individuals outwith the university. These include other universities and academic scholars worldwide, not least are its WUN partners, links with government agencies, funding agencies, media and other quasi and non-government organisations worldwide.

Leeds university sub-organisations are structured hierarchically into administrative units, the largest of which are the nine main subject faculties and the central organisations that deal with generic administration, information systems and services (including the libraries and shared teaching resources). The faculties help to manage schools, centres and institutes - many of which are interdisciplinary and link across multiple sub-organisations [34]. The structure is web like..

In the last 2 years the university web site has been redeveloped and the data structured in a cleaner way to how it was before [35].

The most important part of any organisation like Leeds are the people that are based there and who interact with it. *Leeds people* is rather more nebulous term though as it includes academic scholars and other staff who are now based elsewhere or retired, alumni and visitors past and present. The benefits of mapping these connections using the social web are slowly being realised as Leeds and the UK government try to develop a clearer picture of the universties impacts in terms of the costs and benefits of research and education [36].

2.1. Types of Data

There are two main classes of data, each of which can be sliced and diced in many ways. There is the data about the university as an organisation, and there is the data about things which are the subjects of academic study and scholarly output. There are of course links between these two main classes, and details of who is studying what and who has what expertise and knowledge is the sort of data that many outside organisations are looking for.

In terms of organisation, Leeds is similar to most other universities and others have developed useful diagrams of university data. A useful mind map representation was developed by Chris Gutteridge based at the University of Southampton [37]. The mind map is linked to from three documents which are similar to this and reveal University of Southampton cranking up its LOD efforts about a year ago [38,39,40].

2.2. Current Data

Following a www.leeds.ac.uk website overhaul, the structure of the data on our website is in very good shape, but there are other sub-organisations serving web content for which the data is not so well structured.

Nearly all academic and research staff have a computer user account with a unique user name that allows them to have university personal web pages [41]. Many have at least one sub-organisation home web page, some have several, and others are only listed on pages or are not recognised on the public facing web pages at all. As people leave the university, rarely are their home web pages modified to reflect this fact and left in-situ as a document of their work at Leeds, or in time changed to point to their web home pages at their new organisations as these become known. All too often, people have their web pages deleted or moved which is a shame.

Similarly, nearly all students get a computer user account with a unique user name that allows them to have university personal web pages [41]. Many do not initialise web pages, but all those with computer user accounts can have them in principle. Any personal web pages they create on the university system tend only to persist for a short time (like their university email address) after they move on.

Over time as individuals leave, their usernames may be recycled and this can be problematic.

Some data for visitors, honorary and emeritus scholars and alumni is linked and the university is trying to make use of Facebook to help develop data about its social web.

2.3. Future Data

This is something of a personal vision for future Leeds data.

All staff and academic and research staff past and present and all new students will be assigned a unique identifier that will be used to link their information. This can be used for both open data and more closed data linkage. Similarly, all alumni that request one can also be assigned a unique identifier for linking purposes.

An RDF/XML document will be developed for each sub-organisation and each individual at the university and these will be populated with details about them. RDF documents will also be generated for facilities and services. In time, the Web pages for these entities will be largely a special rendering of the RDF/XML.

Linking and semantic enhancement of existing web content will be done to maximise its potential use. It is not clear yet how best to do this and the use of microformats is probably appropriate.

The data will be structured using the same names as used by other universities to make it easier to find data across multiple universities.

Leeds will start helping to pave the way for further adoption and development of LOD in the sector.

3. Next steps

3.1. Meetings

Web team meeting on 2011-06-23. Can we move quickly to adopt data.leeds.ac.uk and id.leeds.ac.uk as places to hang everything? Do we have support from the top? What meetings should be arranged? How do we resource the development of Leeds LOD? Should we seek funding? What are the next steps?

3.2. Setting up data.leeds.ac.uk

Is it best practice keep have a separate id.leeds.ac.uk? The University of Southampton went down this route. If we do the same, I envisage the data about identifiers at id.leeds.ac.uk having a very similar structure to data.leeds.ac.uk.

Leeds should aim for the same structure and to use the same names and ontologies as other universities developing LOD (c.f. data.soton.ac.uk and data.open.ac.uk). These may be changing, so it is important that the team at Leeds engages in community development rather than simply copying what is there. Let us look forward to taking the step that is updating the data.ac.uk (http://data-ac-uk.ecs.soton.ac.uk/) spreadsheet adding our local representative contact details once it is decided who will undertake this role.

3.3. Promoting openness at Leeds

Openness, transparency and honesty are all part of our integrity at Leeds as stipulated in Our Values of Academic Excellence [42].

Not all our data should be open, but it should probably all be linked. There is a need to keep a considerable amount of university data secured and only accessible in a controlled manner. Role based and individual access control is critically important. However, when thinking about data and whether it should be open, the default should perhaps be that without good reason for keeping the data secure, it should be made available as LOD.

One key benefit of making data openly available as linked data is that it encourages more of the same and for an organisation like Leeds, this can only be a good thing.

3.4. Getting support from the top

In private correspondence with others developing LOD at their universities it has been made clear that they value having “support from the top”. This is not least for resourcing the effort, but moreover for enabling and enhancing collaboration between different departments in the university that develop particular data resources that are to be linked. Interfacing with these sub-organisations will be key and having support of senior managers is likely to be key.

3.5. Collaboration and research

How should we look to work with others in the data.ac.uk (http://data-ac-uk.ecs.soton.ac.uk/) and the broader LOD community? As a start this document was written in the open even though it was not open for community edits. Now it exists, hopefully others can adapt and make use of

it.

3.5.1. Develop unique identifiers use case

A focus on people and organisations with scenario worked examples.

3.5.2. Course finding

Liaison between Tony Hirst, Chris Gutteridge and Paul Santon on the “Killer Coursefinder App” is hopefully happening (c.f. [43,44,45]).

3.5.3. The 3D interactive - community developed - open geographic map

Well maybe we could get around to this. Leeds is a particularly 3D campus with our links on Level 10 and all...

4. Considerata

4.1. Licenses

The current wisdom is that adding an appropriate license for open data encourages its use. Much data about a university will want to be aggregated and used by other organisations and this can be encouraged with an license that does not impose any conditions on the use and further distribution of the data. Should Leeds default to an intentional no license? Again colleagues at the University of Southampton have been considering this [46]. The Open Knowledge Foundation also provide a document for guidance on open data licensing [47].

4.2. schema.org

Schema.org [48] is a joint effort, in the spirit of sitemaps.org [49], to improve the web by creating a structured data markup schema supported by major search engines. On-page markup helps search engines understand the information on web pages and provide richer search results. A shared markup vocabulary makes easier for webmasters to decide on a markup schema and get the maximum benefit for their efforts. Search engines want to make it easier for people to find relevant information on the web. Markup can also enable new tools and applications that make use of the structure.

5. References

1. http://data.gov.uk/linked-data
2. http://linkeddata.org/
3. http://openorg.ecs.soton.ac.uk/wiki/Linked_Data_Guide_for_Newbies
4. http://www.w3.org/DesignIssues/LinkedData.html
5. http://en.wikipedia.org/wiki/Open_data
6. http://opendatamanual.org/
7. http://opendatamanual.org/what-is-open-data/what-is-open-data.html
8. http://www.opendefinition.org/okd/
9. http://linkingyou.blogs.lincoln.ac.uk/tag/linked-data/
10. http://www.w3.org/wiki/WebID

11. http://geohash.org/site/tips.html
12. http://data.ac.uk (http://data-ac-uk.ecs.soton.ac.uk/)
13. http://data.open.ac.uk
14. http://data.soton.ac.uk
15. http://data.online.lincoln.ac.uk
16. http://data.ox.ac.uk
17. http://data.kcl.ac.uk/
18. http://data.ed.ac.uk
19. http://data.gov.uk
20. http://data.gov
21. http://data.gov.uk/organogram
22. http://www.oegov.org/
23. http://data.clarosnet.org/
24. http://blogs.ukoln.ac.uk/locah/about/
25. http://data.archiveshub.ac.uk/
26. http://copac.ac.uk/
27. http://copac.ac.uk/libraries/leeds.html
28. http://europeana.eu
29. http://data.europeana.eu
30. http://linkedgeodata.org
31. http://www.ordnancesurvey.co.uk/oswebsite/opendata/linkeddata.html
32. http://data.ordnancesurvey.co.uk
33. http://www.leeds.ac.uk
34. http://www.leeds.ac.uk/faculty
35. http://coffeepowered.co.uk/2011/06/building-a-killer-coursefinder/
36. http://data.gov.uk/search/apachesolr_search/universities?filters=ss_cck_field_publisher%3A%22Department%20for%20Business%2C%20Innovation%20and%20Skills%22&retain-filters=1
37. http://mind42.com/pub/mindmap?mid=605c3bad-3980-4d4b-9155-75b33af8860d
38. http://blogs.ecs.soton.ac.uk/webteam/2010/07/14/linked-data-for-universities/
39. http://blog.okfn.org/2010/07/20/opening-up-university-infrastructure-data/
40. http://blogs.ecs.soton.ac.uk/webteam/2010/07/17/ecs-infrastructure-rdf-in-the-public-domain/
41. http://www.personal.leeds.ac.uk/
42. http://www.leeds.ac.uk/comms/strategy/strategy/values/index.html
43. http://blog.ouseful.info/2011/03/24/tso-openup-competition-opening-up-ucas-data/
44. http://www.coursedetective.co.uk
45. http://coffeepowered.co.uk/2011/06/building-a-killer-coursefinder/
46. http://blogs.ecs.soton.ac.uk/data/2011/06/14/licenses-in-data-southampton/
47. http://wiki.okfn.org/OpenDataLicensing
48. http://schema.org/
49. http://www.sitemaps.org/
50. Recursive references
    a. http://t.co/fjAJbyr

b. https://docs.google.com/document/d/1EgzYQKK1iq5SMS1Hn1hOfVO7yvJrGYESpxZCBPIOmwc/edit?hl=en_US