Linked Data – The Future for Open Repositories?

8th June 2011
Open Repositories 2011, Austin, Texas, USA

Adrian Stevenson
LOCAH Project Manager
“The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web.”

“the Semantic Web is the goal or end result… Linked Data provides the means to reach that goal”

From ‘Linked Data: The Story So Far’ - Heath, Bizer and Berners-Lee 2009
The goal of Linked Data is to enable people to share structured data on the Web as easily as they can share documents today.

Bizer/Cyganiak/Heath Linked Data Tutorial, linkeddata.org
• But haven’t we been putting linked data on the web for years?
  – In CSV, relational databases, XML etc?
• Well yes, but these approaches are not so easy to integrate
• Web 2.0 mashups work against a fixed set of data sources
• Linked Data applications operate on top of an unbound, global data space.
Data.gov.uk

Officially launched 21st January 2010

Opening up government

Facts, figures, apps and more
Keith Richards

Mick Jagger, Keith Richards And Ron Wood

Biography

Keith Richards (born 18 December 1943) is an English musician, co-songwriter, and founding member of the rock band the Rolling Stones. His innovative rhythm guitar playing was cited when Rolling Stone magazine ranked Richards as the "10th greatest guitarist of all time." Fourteen songs Richards wrote with songwriting partner and the Rolling Stones' vocalist Mick Jagger were listed on Rolling Stone's "500 Greatest Songs of All Time".

Read more at Wikipedia...
Linked Data

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.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

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.
URIs and HTTP

- A ‘Uniform Resource Identifier’ (URI) provides a simple and extensible means for identifying a resource - RFC 3986
- HTTP URIs may be ‘de-referenced’
  - A URL is a type of URI
- HTTP URIs are used for “real world” things
  - http://adrianstevenson.com/id/me
  - http://dbpedia.org/resource/Love
RDF

• Resource Description Framework
  – a language for representing information about resources on the Web
  – RDF can be used to represent things identified on the Web, even when they cannot be directly retrieved on the Web

• Describes relations using ‘triples’

• http://www.w3.org/TR/REC-rdf-syntax/
Triples

- Triples statements
  - ‘Things’ have ‘properties’ with ‘values’
  - Subject – Predicate - Object

- Triples are the basis of RDF
What is the LOCAH Project?

- **Linked Open Copac and Archives Hub**
- Funded by #JiscEXPO 2/10 ‘Expose’ call
- 1 year project. Started August 2010
- http://blogs.ukoln.ac.uk/locah/
tag: #locah
What are the Archives Hub and Copac?

- Archives Hub is an aggregation of archival descriptions from archive repositories across the UK
  - http://archiveshub.ac.uk
- Copac provides access to the merged library catalogues of libraries throughout the UK, including all national libraries
  - http://copac.ac.uk
What is LOCAH Doing?

- Part 1: Exposing Archives Hub & Copac data as Linked Data
- Part 2: Creating a prototype visualisation
- Part 3: Reporting on opportunities and barriers
LOCAH Linked Data

- If something is identified, it can be linked to
- We can then take items from one dataset and link them to items from other datasets
HTTP URI Patterns

• Need to decide on patterns for URIs
• Following guidance from W3C ‘Cool URIs for the Semantic Web’ and UK Cabinet Office ‘Designing URI Sets for the UK Public Sector’

http://data.archiveshub.ac.uk/id/findingaid/gb1086skinner ‘thing’ URI

… is HTTP 303 ‘See Other’ redirected to …

http://data.archiveshub.ac.uk/doc/findingaid/gb1086skinner document URI

… which is then content negotiated to …

http://data.archiveshub.ac.uk/doc/findingaid/gb1086skinner.html
http://data.archiveshub.ac.uk/doc/findingaid/gb1086skinner.rdf
http://data.archiveshub.ac.uk/doc/findingaid/gb1086skinner.turtle
http://data.archiveshub.ac.uk/doc/findingaid/gb1086skinner.json
Enhancing our data

• Already have some links:
  – Language - lexvo.org URIs for languages of archival materials
  – Time - reference.data.gov.uk URIs for time periods
  – Location - using both UK Postcodes URIs and Ordnance Survey URIs
  – Names - Virtual International Authority File
    • Matches and links widely-used authority files - http://viaf.org/
  – Names - DBPedia

• Also looking at:
  – Subjects - Library Congress Subject Headings and DBPedia
LOCAH
This is the homepage of the LOCAH Linked Archives Hub test dataset.

LOCAH Project
The LOCAH project receives funding from JISC under the JiscEXPO strand of its Information Environment programme.
The main aims of the project are to make available data from the Archives Hub and from Copac in the form of Linked Data.
The participants in the project are Mimas, UKOLN, Eduserv, Talis, and OCLC Research.
The project runs from August 2010 until July 2011.

For more information on LOCAH project activity, see the LOCAH project blog.

LOCAH Linked Archives Hub Test Dataset
This is a small test dataset describing archives held by UK higher education institutions. The data is derived from a sample of the archival finding aids held by the UK Archives Hub.
It is hosted on an instance of the Talis Platform triplestore.
The URIs for resources described in the dataset follow the following pattern:
`http://data.archiveshub.ac.uk/id/` +
Example resources are:
http://data.archiveshub.ac.uk/id/archivalresource/gb1086skinner
http://data.archiveshub.ac.uk/id/person/ncarules/skinnerbeverley1938-1999artist
http://data.archiveshub.ac.uk/id/place/repository/gb1086
http://data.archiveshub.ac.uk/id/concept/unesco/womenartists

The data references terms from (amongst others) the following RDF vocabularies:
http://purl.org/dc/terms/
http://xmlns.com/foaf/0.1/
http://www.w3.org/2004/02/skos/core#
http://www.openarchives.org/ore/terms/
http://linkedevents.org/ontology/
http://data.archiveshub.ac.uk/def/

The SPARQL endpoint for the dataset is: http://data.archiveshub.ac.uk/sparql.
An RDF dump of the dataset is available: http://data.archiveshub.ac.uk/dump/linkedarchiveshub.zip.

Please note: at this time, the dataset should be considered unstable, and its content may change substantially at any time without warning.
The LOCAH Linked Archives Hub data and content is licensed under a Creative Commons CCO 1.0 licence.
Archives repository postcode data kindly supplied by The National Archives Archon Directory.
How are we creating the Visualisation Prototype?

• Based on researcher use cases
• Data queried from Sparql endpoint
• Use tools such as Simile, Many Eyes, Google Charts
• Also looking at custom built prototype
Local Timemap Visualization

Locah project data for travel and exploration

Mungo Park
Name: Mungo Park
Occupation: African explorer
Size of archive: 4 boxes
Dates of archive: 1818 - 1828
School of Oriental and African Studies
Browse Hub Linked Data http://data.archiveshub.ac.uk/id/archivalresource/gb102ms282539
Data Modelling

- Complexity
  - Archival description is hierarchical and multi-level
- Dirty Data

Licensing

- ‘Ownership’ of data
- Hard to track attribution
- CC0 for Archives Hub data
- Copac license decision in progress
Sustainability

• Can you rely on data sources long-term?
• http://lcsh.info

Provenance

• data ‘watermarked’

<http://data.archiveshub.ac.uk/doc/archivalresource/gb1086skinner> rdf:type foaf:Document

• Scaling issues
Future for Open Repositories?

- Repository data can ‘work harder’
- New channels into your data
- New connections with other data sources
- Researchers are more likely to discover sources
- ‘Hidden' collections of repositories become of the Web
Attribution and CC License

- Sections of this presentation adapted from materials created by other members of the LOCAH Project
- This presentation available under creative commons Non Commercial-Share Alike:

http://creativecommons.org/licenses/by-nc/2.0/uk/