

# Linked Open Data: a new resource for eResearch

Anne M. Cregan<sup>1, 2</sup>

<sup>1</sup>Intersect Australia Ltd, Sydney, Australia, anne.cregan@intersect.org.au

<sup>2</sup>National ICT Australia (NICTA), Sydney, Australia, anne.cregan@nicta.com.au

## INTRODUCTION

The Open Data Movement aims to make data freely available to everyone. Under Creative Commons License, a Data Commons is rapidly emerging, and the World Wide Web is fast becoming a space not only for linking documents and web pages, but for interlinking data sets. This interlinkage is taking place not just at the level of linking a whole data set to other related data sets, but is happening at the level of the individual data items, where individual items within the data set are interlinked to related data items in other data sets. A large component of research data is suitable for linking into the open data cloud, and international researchers have commenced the process of publishing their data sets online as a collaborative research initiative, as it is an excellent way to expose, share, and connect pieces of research data. Greater visibility and ability to process data with a common theme generated by different research groups enables new research studies and insights to emerge.

## THE LINKED OPEN DATA CLOUD

The Linking Open Data Project [1,2] is a community project of the World Wide Web Consortium's Semantic Web Education and Outreach Group (W3C SWEO). The goal of the project is to extend the Web with a data commons by publishing various open data sets on the Web, and making links between data items from different data sources. Since inception in June 2007, the size of the cloud has rapidly exploded and already includes a large variety of open data sets including several research and medical data sets. Figure 1 shows the data sets published and interlinked so far: as at May 2009, the data sets consisted of over 4.7 billion RDF triples interlinked by around 142 million RDF links.

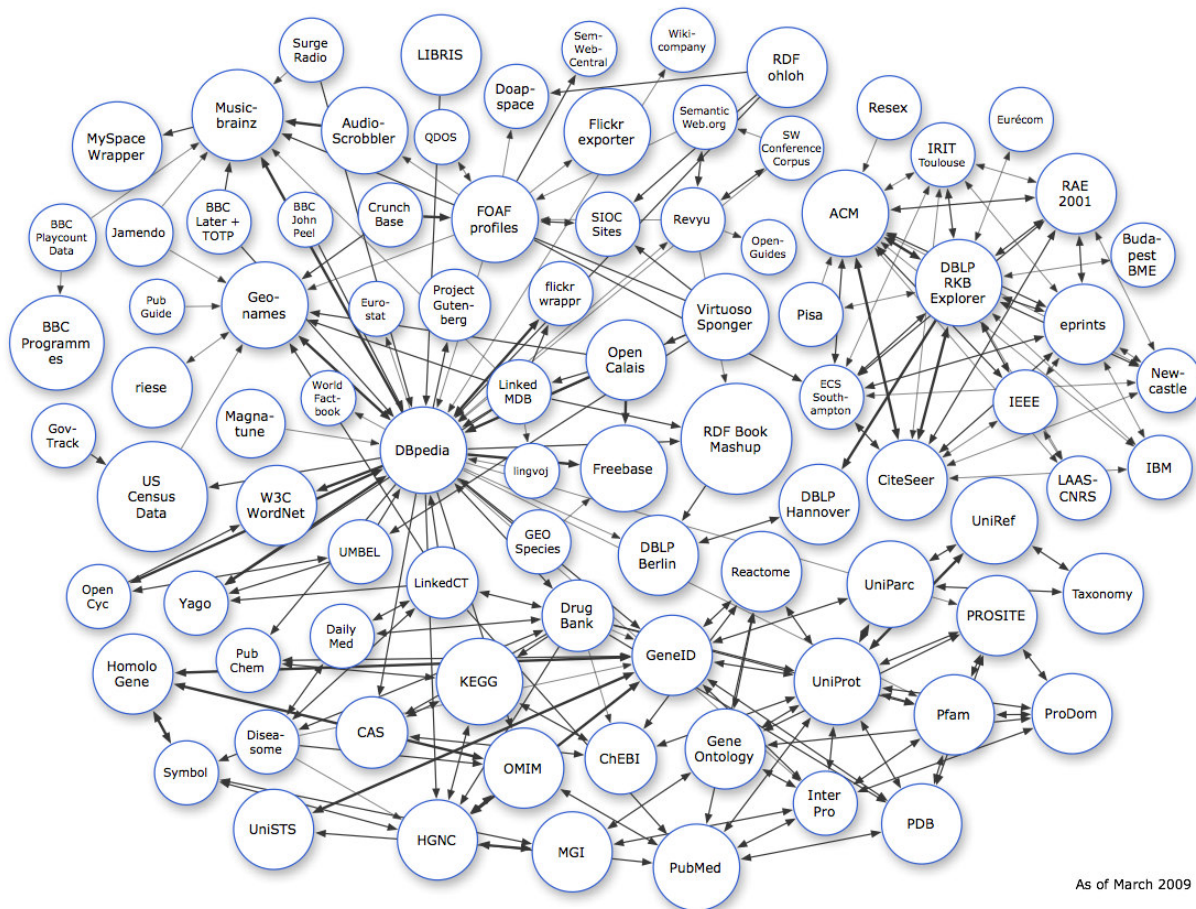


Figure 1: Linked Open Data Cloud data sets as at March 2009

## THE LINKING OPEN DATA PROJECT

The Linking Open Data Project is a community project of the World Wide Web Consortium's Semantic Web Education and Outreach Group (W3C SWEO). The goal of the project is to extend the Web with a data commons by publishing various open data sets in Resource Description Framework (RDF) [3] on the Web, and by setting RDF links between data items from different data sources. Because the LOD cloud is in a machine-processable format, it is possible to "mashup" data sets from different sources to generate new information and insights on an unprecedented scale.

## W3C PRINCIPLES FOR PUBLISHING LINKED DATA

The initiative uses the W3C's recommended standards for data publishing: Uniform Resource Indicators (URIs) to identify data items, and the W3C's Resource Description Framework (RDF) as the mechanism for publishing and linking data. The principles for publishing linked data on the web using these standards are straightforward and simple [4, 5]:

1. **Use URIs as names for things.** Every data item should have its own URI so it can be identified. URIs are the basic building blocks for the web – a web page's URL is a kind of URI.
2. **Make the URIs dereferenceable** so that when someone looks up the URI they can obtain the data item. When you look up a URL, you get a web page; when you look up a URI for a data item, you should get the data.
3. **Describe the data using the RDF data model.**
4. **Publish the data in RDF format.** RDF is used to construct triples which glue URIs together. By using RDF as the publishing format, the data forms a graph where the nodes are common URIs.
5. **Create links between URIs.** Sometimes data sets contain elements which have the same meaning but use different identifiers eg LastName and Surname. Using RDF you can say that one URI is the same thing as another URI with a different name.

## BENEFITS OF USING RDF FOR PUBLISHING DATA

Using the publishing principles, a robust interlinked graph of data is created. This graph is separate from the applications used to display and process it, so it can be made available to any application and furthermore, is fully machine processable. Data items linked using RDF links enable navigation from a data item within one data source to related data items within other sources, using a Semantic Web browser such as Tabulator, Disco or Zitgist [6]. The linked data sets can be queried using the SPARQL query language [7], which is similar to SQL but able to be used over the graph of data formed by RDF triples. As the query results return structured data and not just links to HTML pages, they can be used within other applications.

RDF links can also be followed by the crawlers of Semantic Web search engines, which may provide sophisticated search and query capabilities over crawled data. It is now possible to embed RDF into an XHTML web page using the new RDFa standard [8], which is supported by search engines Yahoo! and Google.

## VALUE TO ERESEARCH COMMUNITY

Publishing research data to the Linked Open Data Cloud has huge potential for researchers in the following ways:

- Make research data available publicly in a reusable format, separated from any particular application.
- Link your research data to data sets produced by other research groups, or data from related areas. For instance, a group of chemists working on a particular compound could link their data to other groups working on the same compound, to data about other compounds and to data produced by a group of physicists working in a related area from a different angle.
- Navigate, search and query over the linked data. The linked data can be used in a way similar to data in a relational database, but instead of being limited to your own data sets you can access all the linked data.
- If desired, there is potential to use the Semantic Web standards to derive new information using formal reasoning.

## REFERENCES

1. *Linked Data Website* Available from <http://linkeddata.org/> accessed 29 June 2009.
2. *W3C Linking Open Data Wiki* Available from <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenDataLinking> accessed 29 June 2009.
3. *RDF Language (W3C Recommendation)*. 2004. Available from <http://www.w3.org/RDF/>
4. Berners-Lee, T. *Linked Data*. 2006. Available from <http://www.w3.org/DesignIssues/LinkedData.html>
5. Bizer, C., Cyganiak, R., Heath, T. *How to Publish Linked Data on the Web*. 2007. Available from <http://www4.wiwi.fu-berlin.de/bizer/pub/LinkedDataTutorial/>
6. *Tabulator, Disco and Zitgist RDF Browsers*. Available from: <http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/SemWebClients> accessed 29 June 2009.
7. *SPARQL Query Language for RDF*. 2008. Available from <http://www.w3.org/TR/rdf-sparql-query/>
8. *RDFa Primer*. 2008. Available from <http://www.w3.org/TR/xhtml1-rdfa-primer/>