7 things you should know about…Linked Data

Scenario

The traditional approach of sharing data within silos seems to have reached its end with Web advancing to an era of opening data. From governments and international organizations to local cities and institutions, there is a widespread effort of opening up and interlinking data. Two important concepts have been coined in this context:

- **Open Data,** defined as "*data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and share alike*"[1]; and
- **Linked Data,** associated to the technical interoperability of data, which enables to connect data from a variety of sources (related to the Semantic Web architecture)[2].

While Open Data refers to data freely available without restrictions[3], Linked Data is refereeing to machine-readable data and semantically linked. Therefore data can be open but not linked or linked but not open, however if data is open and linked it then becomes Linked Open Data.

The main difference between the web of hypertext and the Semantic Web is that while the first links html pages or documents, the second goes beyond the concept of document and links structured data. In this context, Linked Data is the set of best practices for publishing and connecting structured data on the Web.

This particular scenario is beneficial for digital repositories, as a way to enhance the visibility and interoperability of data by linking their content into the wider Web of Data.

1. What is Linked Data and Linked Open Data?

Linked Data refers to a set of best practices for publishing, sharing, and interlinking structured data on the Web. Its main objective is to liberate data from silos that are framed by proprietary database schemas following four rules, defined by Tim Berners-Lee in 2006[4], as follows:

1. Use of Uniform Resource Identifiers (URIs) for identifying entities or concepts uniquely in the world
2. Use of HTTP URIs for retrieving resources or descriptions of resources
3. Use of standard formats like RDF for structuring and linking descriptions of things
4. Use of links to other related URIs in the exposed data to improve discovery of related information on the Web

These principles are defined as rules[5], but in reality they are rather recommendations or best practices for the development of the Semantic Web. Data can be published meeting only the first three. However, the failure to achieve the fourth rule is what makes data less visible and, therefore, less *sharable*, *extensible* and *re-usable*.

Linked Open Data (LOD) is Linked Data distributed under an open license that allows its reuse for free. In 2010, Tim Berners-Lee defined a 5-star rating scheme (http://www.w3.org/DesignIssues/LinkedData)[6] to encourage data providers to provide linked data under open licenses. The scheme uses gold stars to evaluate the availability of linked data as linked open data:

★	Data available on the web in any format, even using PDF or image scan, but with an Open licence
★★	Data delivered as machine-readable structured data, e.g. excel instead of image scan of a table
★★★	Data available in a non-proprietary format, e.g. CSV instead of excel
★★★★	All the above plus, data using open standards from W3C, e.g. RDF and SPARQL, to identify things and properties, so that people can point at other data
★★★★★	All the above, plus, to link data to other people's data to provide context

2. How does it work?

In order to link data distributed across the Web, a mechanism is needed to specify the meaning of connections between items described in the data. This standard mechanism is RDF, the Resource Description Framework for metadata on the Web developed by the W3C (http://www.w3.org/)[7].

It is based on the idea of declaring resources using the expression in the form subject-predicate-object. This expression is known as RDF triple. An RDF triple contains three components, all with its own URI:

- Subject, a URI, a person, or node, is the entity to which we refer;
- Predicate is the property or relationship you want to set about the subject;
- Object is the value of the property or another resource that establishes the relationship.

By using URIs to link data, the Web becomes a kind of large database that allows people and machines to explore the information referenced and interconnected. The Web-based on Linked

Data is a breakthrough in content syndication, which uses external data sources to create new services[8].

Simply transforming database schemas into RDF does not create Linked Open Data. There is a chance to get stuck at the 4th star in the 5-star rating scheme. To avoid creating RDF silos, it is necessary to create automatic links between RDF triple stores on the web. The easiest way to facilitate the establishing of automatic linking between datasets is the use of standard vocabularies, including standard vocabularies for describing data or metadata elements and standard vocabularies for indicating values.

3. Who is doing it?

International initiatives promoting Open Data and Linked Data

In the context of Open Data, sponsors from the European Commission, the U.S. Government, and the Australian Government and other players with the data community launched the Research Data Alliance (RDA)[9] in Gothenburg (Sweden) on Mach 2013. This initiative aims to facilitate the global research data sharing and exchange by the harmonization of data standards and practices. RDA is organised into working and interest groups and plenary meetings are held quarterly; participation from governments, researchers, and practitioners, however activities are open to all interested persons.

The Open Knowledge Foundation (OKF)[10] is a non-profit organisation dedicated to promoting open data with an extensive experience in building tools and communities. The CKAN, open source data portal platform and Data Hub, a community-run catalogue of datasets available on the Web are part of the projects being managed and promoted by the OKF's staff and communities.

In December 2012, the Open Data Institute (ODI)[11] was launched in the UK with the objective to promote new business and culture around open data by creating economic, environmental, and social value and by promoting standards. The Institute was founded (http://www.wired.co.uk/news/archive/2012-05/21/open-data-institute-plans) by Tim Berners-Lee and Nigel Shadbolt with funding from the UK Government and Omidyar Network. ODI has recently launched the Open Data Certificates[12] to help to find, understand and use published open data. The objective is to create mechanisms to bring accuracy to the publication, dissemination and usage of open data according to the needs of business, governments, and citizens.

At the Open Government Partnership Summit in London on October 2013, the Global Open Data for Agriculture and Nutrition (GODAN)[13] was launched to support global efforts to make agricultural and nutritionally relevant data available, accessible, and usable for unrestricted use worldwide. In the same context, and since 2008, the CIARD[14] Movement works to expand

openness by fostering collaborative approaches and mutual learning towards open agricultural knowledge for development.

Accessing Linked Open Data Sets

Datahub.io[15] is the data management platform provided by OKF (http://okfn.org/services/) to publish, register or share datasets. The web interface is a way to help people find and search published datasets. It is also possible to manage groups of datasets, e.g. the Linking Open Data Cloud diagram uses the descriptions of the data sets from the group *Linking Open Data Cloud*.

The Linking Open Data Cloud[16] diagram (Figure 1) shows datasets that have been published in Linked Data format by contributors from the Linking Open Data community project and other individuals and organisations. In order to be present in the graph, data sources should publish data as follows:

- resolvable *http:// (or https://) URIs*
- resolvable to *RDF data* in any standard RDF format. e.g. RDFa, RDF/XML, Turtle, N-Triples
- containing at least 1,000 triples
- connecting via RDF with links to at least one dataset already in the diagram (it is required at least 50 links)
- being accessible the entire dataset via RDF crawling, via an RDF dump, or via a SPARQL endpoint

(http://lod-cloud.net/)
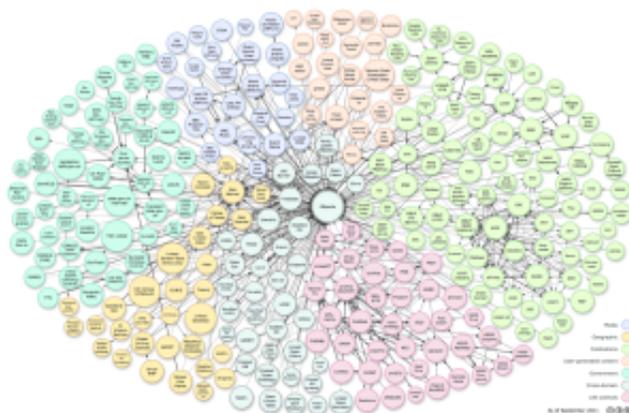
Last updated of the Linking Open Data cloud diagram in
2011 by Richard Cyganiak and Anja Jentzsch. http://lod-
cloud.net/

Another dynamic force graph version of the LOD Cloud is the **Linked Open Data Graph**[17] which highlights the ratings of datasets using a Protovis[18] version of the Linked Open Data Cloud using data made available by the CKAN API.

4. Why is it significant?

If all the data on the Web were open and linked, it would be easier to establish information systems combining different distributed data repositories. Thus, the Web of Data would enable access and sharing of data and knowledge without barriers.

5. What are the downsides?

The quantity of published Linked Data increases day by day. However the fact that some of the data available might be either irregularly updated, or already available in other formats and APIs might become an issue. This is not happening will all the datasets, but it needs to be taken under consideration. Additionally, more data needs to be available to *share*, *extent* and *re-use*. Data should be urgently published as Linked Data on the Web with appropriate licenses and provenance information. Without data to be linked to there is a risk of creating RDF silos. There is also a lack of applications and tools to exploit Linked Data. Existing open issues make the development of Linked Data based applications a challenge, due to the difficulties to integrate data in different formats and from multiple sources, the discovery of data or the usability of user interfaces.

6. Where is it going?

The proposal of the Semantic Web as a common framework that allows data to be shared and reused across application, enterprise, and community boundaries was already launched in 2001. However, its practical application was not possible until governments and research organizations started to discuss and promote the publication of open data worldwide. Institutions will continue taking steps along the road to liberating government and research data, with the objective to support global efforts to make data available, accessible, and usable for unrestricted use worldwide.

7. What are the implications for institutional repositories?

To publish open access documents on the Web is not enough for being part of the Web of Data. Different development stages, internal data structures, and reality of their practices may jeopardize the dissemination and accessibility of the open access documents. Existing methodologies, standards and technologies available to facilitate the publication and exchange of data should be much more accessible to information management specialists.

There are several benefits for institutional repositories in providing access and visibility to the scientific production on the Web when consuming and publishing Linked Open Data:

- Opportunity to develop local and wider services on open access resources aggregating

additional information resources. Different types of information like bibliographic resources, statistics or geospatial information could be mashed-up and displayed in a single interface.

- Enrichment of data from other Linked Data sources, especially controlled vocabularies, authority data and syntax encoding standards. Traditional institutional repository software should facilitate the integration of vocabularies published as Linked Open Data.
- Increased exposure of institutional repository collection to web search engines.
- Collections easier to access while also making new applications more useful.
- Reduction of redundancy of bibliographic descriptions on the Web.

The W3C Library Linked Data Incubator Group[32] (2010-2011) mentioned in its recommendations to encourage libraries to participate in the Linked Data framework:

> *"the web of information should be embraced, both by making data available for use as Linked Data and by using the web of data in information services. Ideally, data should integrate fully with other resources on the Web (…) In engaging with the web of Linked Data, libraries can take on a leadership role grounded in their traditional activities: management of resources for current use and long term preservation; description of resources on the basis of agreed rules; and responding to the needs of information seekers"[33].*

In an ideal world, all data would be linked on the Web. This would establish information systems combining different data from distributed repositories. A scenario like this is not science fiction.

Special thanks to Imma Subirats from FAO of the UN for creating this section.

Imma Subirats has been working as senior knowledge and information management officer at the Food and Agriculture Organization of the United Nations (FAO) since 2006. She works on the advise of standards, tools and good practices for the management and exchange of data to academic, research, private and governmental institutions worldwide. She is also actively promoting open access and open data in the agricultural research context. In recent years, she has been working on the facilitation of the AIMS community and portal, space for accessing and discussing information management standards, tools and methodologies with the objective to connect information specialists worldwide to support the implementation of structured and linked information and knowledge.
Additionally, in 2003 she co-founded e-LIS, e-prints in Library and Information Science, a voluntary enterprise with the objective of creating and maintaining a platform where professionals in Library and Information Science can exchange and access research publications and data. e-LIS is possible thanks to the voluntary efforts of 60 librarians from more than 30 countries and provides services to 17,000 authors in Library and Information Science.

[1] Open Knowledge Foundation. *Open Definition*. Available at http://opendefinition.org (http://opendefinition.org), Accessed December 11 2013.

[2] Baker, Tom et al., 2011. *Library Linked Data Incubator Group Final Report*. Available at http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/ (http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/), Accessed December 11 2013.

[3] Wikipedia. *Open Data*. Available at http://en.wikipedia.org/wiki/Open_data (http://en.wikipedia.org/wiki/Open_data), Accessed December 11 2013.

[4] Berners-Lee, Tim, 2009. *Linked Data*. Available at http://www.w3.org/DesignIssues/LinkedData (http://www.w3.org/DesignIssues/LinkedData), Accessed December 11 2013.

[5] Berners-Lee, Tim, 2009. *The Four Rules*. Available at http://www.w3.org/DesignIssues/LinkedData (http://www.w3.org/DesignIssues/LinkedData), Accessed December 11 2013.

[6] Berners-Lee, Tim, 2009. *Is your Linked Open Data 5 Star?*. Available at http://www.w3.org/DesignIssues/LinkedData (http://www.w3.org/DesignIssues/LinkedData), Accessed December 11 2013.

[7] Heath, Tom; Bizer, Christian. *Linked Data: Evolving the Web into a Global Data*. Available at Spacehttp://linkeddatabook.com/editions/1.0/ (http://linkeddatabook.com/editions/1.0/), Accessed December 11 2013.

[8] FAO of the United Nations. *Agricultural Information Management Standards*. Available at http://aims.fao.org, Accessed December 11 2013.

[9] *Research Data Alliance : Research Data Sharing without Barriers*. Available at https://rd-alliance.org (https://rd-alliance.org/), Accessed December 11 2013.

[10] *Open Knowledge Foundation*. Available at http://okfn.org (http://okfn.org/), Accessed December 11 2013.

[11] *Open Data Institute.* Available at http://theodi.org (http://theodi.org/), Accessed December 11 2013.

[12] *Open Data Certificates*. Available at https://certificates.theodi.org (https://certificates.theodi.org), Accessed December 11 2013.

[13] *Global Open Data for Agriculture and Nutrition*. Available at http://godan.info/ (http://godan.info/) , Accessed December 11 2013.

[14] *CIARD*. Available at http://www.ciard.net (http://www.ciard.net) , Accessed December 11 2013.

[15] Open Knowledge Foundation. *Datahub*. Available at http://datahub.io (http://datahub.io/), Accessed December 11 2013.

[16] Open Knowledge Foundation. *Datahub: Linking Open Data Cloud Group*. Available at http://datahub.io/group/about/lodcloud (http://datahub.io/group/about/lodcloud), Accessed December 11 2013.

[17] *Linked Open Data Graph.* Available at http://inkdroid.org/lod-graph/ (http://inkdroid.org/lod-graph/), Accessed December 11 2013.

[18] Protovis : a graphical approach to visualization. Available at http://mbostock.github.io/protovis/ (http://mbostock.github.io/protovis/), Accessed December 11 2013.

[19] Agricultural Information Management Standards. Available at http://aims.fao.org/ (http://aims.fao.org/), Accessed December 11 2013.

[20] FAO of the United Nations. *AGROVOC*. Available at http://aims.fao.org/standards/agrovoc/ (http://aims.fao.org/standards/agrovoc/), Accessed December 11 2013.

[21] Subirats, Imma and Zeng, Marcia Lei, 2012. *Meaningful Bibliographic Metadata (M2B): Recommendations of a set of metadata properties and encoding vocabularies*. Available at http://aims.fao.org/advice/metadata-beta-version (http://aims.fao.org/advice/metadata-beta-version), Accessed December 11 2013.

[22] *Purpose of the LODE-BD Recommendations*. Available at http://aims.fao.org/lode/bd-2/about (http://aims.fao.org/lode/bd-2/about), Accessed December 11 2013.

[23] Subirats, Imma and Zeng, Marcia Lei, 2012. *LODE-BD Recommendations 2.0: How to select appropriate encoding strategies for producingLinked Open Data (LOD)-enabled bibliographic data*. Available at http://aims.fao.org/lode/bd (http://aims.fao.org/lode/bd), Accessed December 11 2013.

[24] Zeng, Marcia; Subirats, Imma, 2013. *How to select appropriate encoding strategies for producing LOD-enabled bibliographic data* [Webinar] Available at http://aims.fao.org/community/general-information/blogs/webinaraims-how-select-appropriate-encoding-strategies-producing (http://aims.fao.org/community/general-information/blogs/webinaraims-how-select-appropriate-encoding-strategies-producing) Accessed December 11 2013.

[25] FAO of the United Nations. *AgriDrupal*. Available at http://aims.fao.org/tools/agridrupal (http://aims.fao.org/tools/agridrupal), Accessed December 11 2013.

[26] FAO of the United Nations. *AgriOcean DSpace*. Available at http://aims.fao.org/agriocean-dspace (http://aims.fao.org/agriocean-dspace), Accessed December 11 2013.

[27] GFAR. *CIARD RING*. Available at http://ring.ciard.net/ (http://ring.ciard.net/), Accessed December 11 2013.

[28] *AGRIS: International Information System for the Agricultural science and technology*. Available at http://agris.fao.org/agris-search/index.do (http://agris.fao.org/agris-search/index.do), Accessed December 11 2013.

[29] *AGRIS: How it works*. Available at http://agris.fao.org/content/how-it-works (http://agris.fao.org/content/how-it-works), Accessed December 11 2013.

[30] Celli, Fabrizio; Keizer, Johannes, 2013. *Release of AGRIS 2.0: Searching agricultural bibliographic data* [Webinar] Available at http://aims.fao.org/community/blogs/new-webinaraims-release-agris-20-searching-agricultural-bibliografic-data-interlinke (http://aims.fao.org/community/blogs/new-webinaraims-release-agris-20-searching-agricultural-bibliografic-data-interlinke), Accessed December 11 2013.

[32] *W3C Library Linked Data Incubator Group*. Available at http://www.w3.org/2005/Incubator/lld/ (http://www.w3.org/2005/Incubator/lld/), Accessed December 11 2013.

[33] Baker, Tom et al., 2011.