Seven 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. 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

[1] Open Knowledge Foundation. *Open Definition*. Available at 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/, Accessed December 11 2013.
[3] Wikipedia. *Open Data*. Available at 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, Accessed December 11 2013.

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[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[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

[5] Berners-Lee, Tim, 2009. *The Four Rules*. Available at 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, 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/, Accessed December 11 2013.
[8] FAO of the United Nations. *Agricultural Information Management Standards*. Available at http://aims.fao.org, Accessed December 11 2013.

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 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 to publish, register or share datasets. The web interface is a way to help people find and search published datasets. It is also

[9] *Research Data Alliance : Research Data Sharing without Barriers*. Available at https://rd-alliance.org, Accessed December 11 2013.
[10] *Open Knowledge Foundation*. Available at http://okfn.org, Accessed December 11 2013.
[11] *Open Data Institute*. Available at http://theodi.org, Accessed December 11 2013.
[12] *Open Data Certificates*. Available at https://certificates.theodi.org, Accessed December 11 2013.
[13] *Global Open Data for Agriculture and Nutrition*. Available at http://godan.info/ , Accessed December 11 2013.
[14] *CIARD*. Available at http://www.ciard.net , Accessed December 11 2013.
[15] Open Knowledge Foundation. *Datahub*. Available at http://datahub.io, Accessed December 11 2013.

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

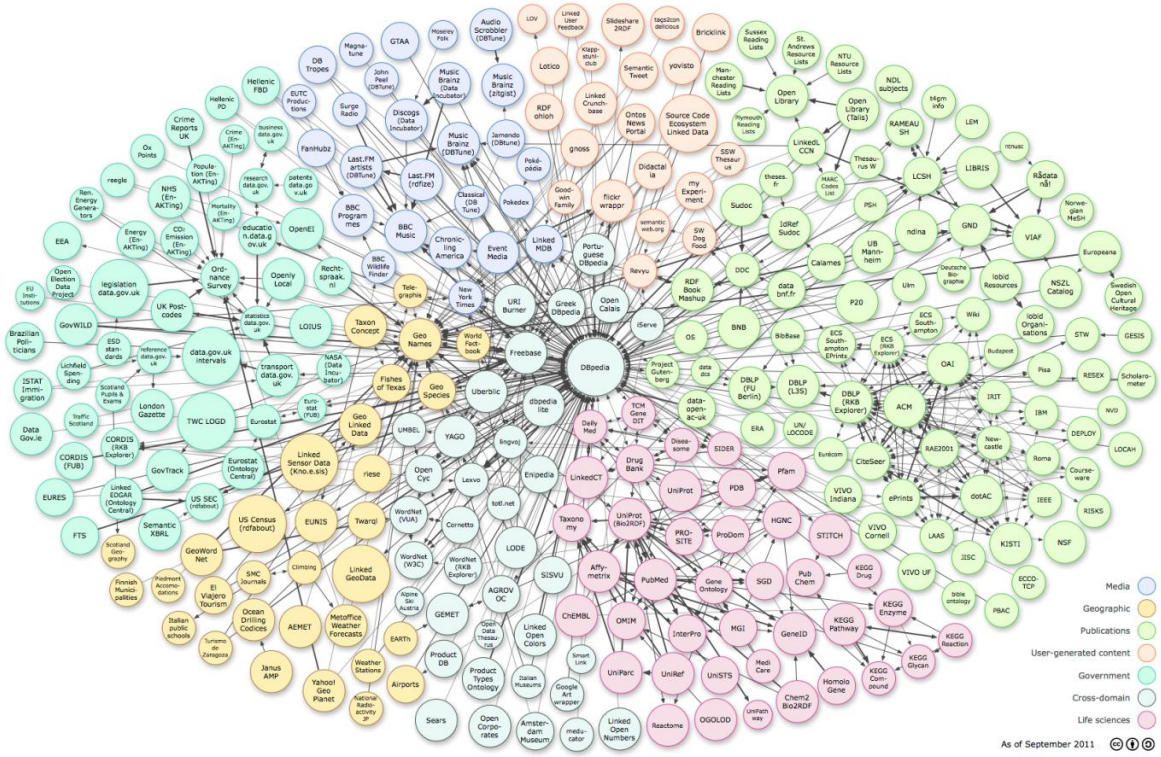

Figure 1. Last updated of the Linking Open Data cloud diagram in 2011
by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

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.

[16] Open Knowledge Foundation. *Datahub: Linking Open Data Cloud Group*. Available at http://datahub.io/group/about/lodcloud, Accessed December 11 2013.
[17] *Linked Open Data Graph.* Available at http://inkdroid.org/lod-graph/, Accessed December 11 2013.
[18] Protovis : a graphical approach to visualization. Available at http://mbostock.github.io/protovis/, Accessed December 11 2013.

A Use Case in the Agricultural Domain

In the Linked Open Data paradigm, institutional repositories have the opportunity to enhance *shareability, extensibility,* and *re-usability* of their data by ensuring:

- content stable, discoverable, and readable data by both machines and humans
- use of appropriate well established metadata standards and emerging Linked Open Data enabled vocabularies
- use of controlled vocabularies, authority data and syntax encoding standards in metadata statements
- use of resource URIs as data values when they are available

In the agricultural domain, the Food and Agriculture Organization of the United Nations provides the agricultural information management community with standards, tools and good practices to assist owners of open repositories to take advantage of the new generation of web-based technologies to increase visibility. This work is facilitated through the global community on Agricultural Information Management Standards (AIMS)[19]. On the AIMS portal, the community advocates for and promotes the application and use of semantic technologies and standards, interoperability of agricultural information and systems, and recommendations on managing research outputs.

The AIMS community is actively working on providing support and tools for all the processes needed for the publication and consumption of bibliographic content as Linked Open Data in the agricultural domain:

1 Linking content by using widely-used controlled vocabularies

The use of meaningful metadata for bibliographic content description and the use of shared vocabularies are primary steps in facilitating interoperability. AIMS provides recommendations to facilitate this exchange of data and information sharing by encouraging the use of authority data, controlled vocabularies, and syntax encoding standards.

AGROVOC[20] is a subject vocabulary covering areas that include food, nutrition, agriculture, fisheries, forestry and environment. To date, AGROVOC contains over 32,000 concepts organized in a hierarchy, where each concept may have labels in up to 22 languages. AGROVOC is available as a Linked Open Data, aligned with more than 10 vocabularies. The Linked Data version of AGROVOC is in RDF/SKOS-XL. Data is accessible to machines through a SPARQL endpoint, and to humans by means of a HTML pages generated with Pubby.

2 Selecting appropriate encoding strategies for producing metadata

Recommendations are essential on what standards[21] to follow and how to prepare LOD-ready metadata to be exposed for service providers. A great number of metadata-related standards have been developed during the last two decades by different communities for specific purposes

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

to guide the design, creation, and implementation of data structures, data values, data contents, and data exchanges. This makes a bit difficult the decision on *what standards to use*.

Decisions regarding what standard(s) to adopt directly impact the degree of LOD-readiness of the bibliographic data. To employ well-accepted metadata element sets and value vocabularies has already shown great benefits and potentials in terms of resource discovery, reuse, sharing, and the creation of new content based on Linked Data. However, in the context of producing LOD-enabled bibliographical data, data and service providers are likely to have many specific questions related to the encoding strategies, e.g. *what metadata standard(s) to follow in order to publish bibliographic data as Linked Data? What minimal set of properties a bibliographic dataset needs to include to insure meaningful data sharing? if the controlled vocabulary is available as Linked Data, what kind of values should be exchanges through our repository, the literal form representing a concept or the URI identifying the concept?* [22]

LODE-BD[23] was born in this context with the purpose of assisting data providers in selecting appropriate encoding strategies for producing meaningful **L**inked **O**pen **D**ata (LOD)-**e**nabled **b**ibliographical **d**ata. The LODE-BD recommendations are applicable for structured data describing bibliographic resources such as articles, monographs, theses, conference papers, presentation materials, research reports, learning objects, etc. – in print or electronic format[24].

3 Integrating metadata standards, controlled vocabularies, authority data and syntax encoding standards in repository tools

The promotion of information management standards has shown that providing tools that implement good practices in the creation, management and exchange of metadata is a key factor to success. Providing metadata and vocabularies via Linked Open Data, Web services and file downloads is not enough. The additional customization of information management tools pre-packaged with such standards and services is fundamental to ensure interoperability among information management systems.

Information system customizations based on two open source content and digital repository management systems, Drupal and DSpace, have been created under the umbrella of AIMS. These customizations - AgriDrupal[25] and AgriOceanDSpace[26] - facilitate the publication of interoperable and re-usable metadata that describe agricultural research information.

4 Discovering information services by registering them in directories

Once the decision on using tools that integrate widely used metadata standards and controlled vocabularies is taken and repositories are in place, the discovery of the information services

[22] *Purpose of the LODE-BD Recommendations*. Available at 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, 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 Accessed December 11 2013.
[25] FAO of the United Nations. *AgriDrupal*. Available at 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, Accessed December 11 2013.

hosting Linked Open Data is essential for building aggregators. The CIARD Routemap to Information Nodes and Gateways (RING)[27], a global registry of web-based services, provides a space for information providers to register their services in various categories with the objective to facilitate the discovery of sources of agriculture-related information across the world.

5 Aggregating information from different resources using mash-up web applications

AGRIS[28], one of the most important world-wide information systems in the area of the agricultural sciences, benefits of all the steps described above. AGRIS uses AGROVOC as backbone to index its records and linked for externals resources; aggregates information using the recommendations on metadata standards described on LODE-BD; and consumes data exposed by data providers registered on the CIARD RING. To date, it hosts more than 7 million of bibliographic records.

AGRIS uses Linked Open Data methodologies to link the bibliographic records with other to related datasets on the web with the objective to enrich the information provided in the AGRIS records. AGRIS interlinks with datasets like DBPedia, World Bank, FAO Geopolitical Ontology, Nature OpenSearch, Global Biodiversity Information Facility and Biodiversity International[29], using AGROVOC[30]. More than 180 million triples have been generated so far.

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[31] was already launched in 2001.

[27] GFAR. *CIARD RING*. Available at 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, Accessed December 11 2013.
[29] *AGRIS: How it works*. Available at 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, Accessed December 11 2013.
[31]

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.

[32] *W3C Library Linked Data Incubator Group*. Available at http://www.w3.org/2005/Incubator/lld/, Accessed December 11 2013.
[33] Baker, Tom et al., 2011.