Linked Open Data in Library Use Today

Gustavo Candela

2024-06-04

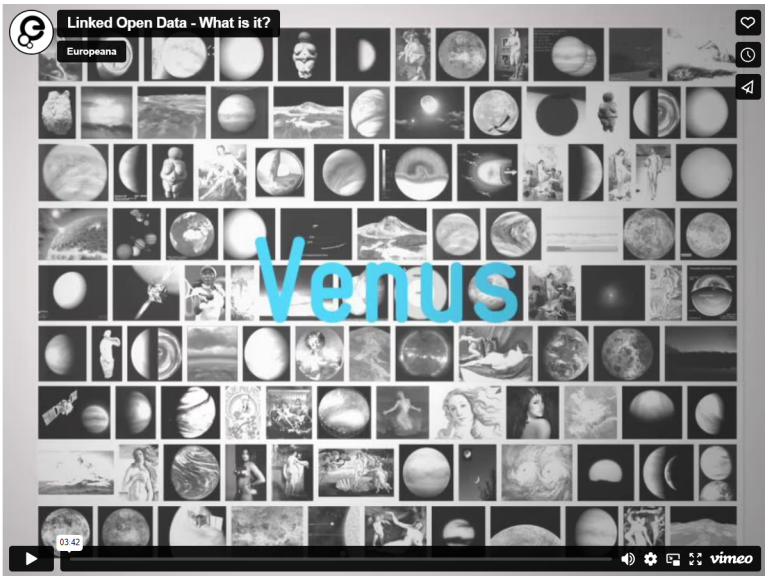
A quick overview of the Collections as Data movement and what it can do for your library, along with some practical advice on how to take steps to transform your own library collections.

Introduction

The Semantic Web was first introduced in the 2000s by Tim Berners Lee as an extension of the current Web. Instead of providing information in the form of *documents* and unstructured text like in traditional webpages, the Semantic Web facilitates the publication of machine-readable *data* on the web through standards such as Resource Description Framework (RDF) and Web Ontology Language (OWL).

What does publishing linked open data enable? Linked Open Data (LOD) is a method of publishing structured data about things using the RDF to enable interlinking and semantic queries across datasets. The data is organised in "triples", each consisting of a subject (e.g., Named Person), predicate (IsAuthorOf), and object (Book Title), identified by Uniform Resource Identifiers (URIs) to ensure global uniqueness and interoperability. It allows metadata to be connected and enriched, so that different representations of the same content can be found, and links made between related resources.

Have a quick look at this video from Europeana explaining the high-level basic principle of LOD before we dive a bit deeper into how it practically works:



Linked Open Data | Europeana PRO

How it works

RDF triples are the fundamental building blocks of Linked Open Data. Triples follow the RDF standard and consist of three components:

1. **Subject**: This is the entity or resource being described. It is usually represented by a URI that uniquely identifies the resource.
2. **Predicate**: This represents the relationship or property of the subject. It is also identified by a URI and specifies the type of relationship between the subject and the object.
3. **Object**: This is the value or resource that is related to the subject. The object can be another URI (representing another resource) or a literal value (such as a string or number), amongst others.

An example of a triple stating "Miguel de Cervantes is author of El Quijote." would look like this:

- **Subject**: <http://www.wikidata.org/entity/Q5682> (Wikidata URI reference to Miguel de Cervantes)
- **Predicate**: <http://purl.org/dc/terms/creator> (Dublin Core URI term for creator/author)
- **Object**: <http://www.wikidata.org/entity/Q480> (Wikidata URI reference to the book El Quijote)

In a 2020 survey of LIBER members, the LIBER Linked Open Data Working Group identified the following as the most frequently used datasets by libraries to enrich their catalogues but

there are many more that can be used depending on your needs Linked Data Survey (oclc.org) and some examples of advanced data models are Bibliographic Framework (BIBFRAME) and Library Reference Model (LRM). In addition, the lod-cloud provides more than one thousand LOD repositories classified by categories and based on different domains such as geography and government.

NAME OF THE DATASET	DESCRIPTION	SOURCE
GeoNames	Contains over 25 million geographical names and consists of over 11 million unique features whereof 4.8 million populated places and 13 million alternate names.	geonames.org
Wikidata	Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.	wikidata.org
DublinCore	The Dublin Core Metadata Initiative supports innovation in metadata design and best practices. DCMI is supported by its members and is a project of ASIS&T.	dublincore.org
Virtual International Authority File (VIAF)	OCLC's Virtual International Authority File (VIAF), an aggregation of over 40 authority files from different countries and regions.	viaf.org
Library of Congress. International Standard Name Identifier (ISNI)	Library of Congress' Linked Data Service with over 50 vocabularies. Although usage fluctuates, it receives 500,000 to a million requests a day.	id.loc.gov/vocabulary/identifiers/isni

So let's go back to our triple that describes the relationship between the resource "Miguel de Cervantes" and the book of "El Quijote". When different triples share the same URI for a subject, predicate, or object, they create a connection. For example:

Triple 1: Miguel de Cervantes is author of El Quijote

- **Subject**: <http://www.wikidata.org/entity/Q5682> (Wikidata identifier for the author Miguel de Cervantes)
- **Predicate**: <http://purl.org/dc/terms/creator> (Dublin Core term for creator/author)
- **Object**: <http://www.wikidata.org/entity/Q480> (Wikidata identifier for the work El Quijote)

Triple 2: El Quijote is a work of Spanish Literature

- **Subject**: <http://www.wikidata.org/entity/Q480> (Wikidata identifier for the work El Quijote)
- **Predicate**: <http://purl.org/dc/terms/subject> (Dublin Core term for subject)
- **Object**: <http://dbpedia.org/resource/Spanish_literature> (DBpedia identifier for Spanish literature)

Here, the object of the first triple (<http://www.wikidata.org/entity/Q480>) is the subject of the second triple, linking information about the book to information about its subject matter. So an example catalogue record combining many triples then might look like:

<http://example.org/catalogue/El_Quijote> >rdf:type schema:Book; schema:name "El_Quijote"; schema:author **<http://www.wikidata.org/entity/Q5682>**; schema:genre <http://dbpedia.org/resource/Novel>; schema:inLanguage <http://id.loc.gov/vocabulary/iso639-1/es>; schema:datePublished "1605"; schema:about <http://dbpedia.org/resource/

Spanish_literature>; schema:about <http://dbpedia.org/resource/Spanish_Golden_Age>; schema:sameAs <http://dbpedia.org/resource/Don_Quixote>.

<http://www.wikidata.org/entity/Q5682> >rdf:type schema:Person; schema:name "Miguel de Cervantes"; schema:birthPlace **<http://dbpedia.org/resource/Alcala_de_Henares>**; schema:birthDate "1547-09-29".

<http://dbpedia.org/resource/Alcala_de_Henares>** >rdf:type schema:Place; schema:name "Alcalá de Henares"; geo:country **<http://sws.geonames.org/2510769/>**.

<http://sws.geonames.org/2510769/> >rdf:type schema:Country; schema:name "Spain" .

Relevance to the Library Sector (Case Studies/Use Cases)

GLAM institutions and in particular, libraries, have played a leading role in the publication of their data, primarily collections metadata, as LOD and using them including:

- Bibliothèque nationale de France
- Biblioteca Virtual Miguel de Cervantes
- British Library
- Europeana
- Library of Congress
- National Library of Scotland
- National Library of Spain

Additional examples from other related domains such as museums and Digital Humanities initiatives are the Rijksmuseum and Smithsonian American Art Museum, and Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH).

The benefits of the publishing and use of the Semantic Web and LOD for:

- **Semantic Enrichment:** LOD helps libraries improve searchability and enables more precise queries by enriching existing catalogue records. Libraries have started to enrich their catalogues with external LOD repositories in order to provide additional contextual information that may be missing from your own catalogue (e.g., author nationalities (VIAF), geographic coordinates (GeoNames) relating to birth places of authors, or related subjects (Library of Congress Subject Headings). As in the example above a catalogue record for the book "El Quijote," could be enriched with metadata about the author, language, publication date, related literary movements, and geographical information, all connected through LOD triples.
- **Interconnectedness:** LOD allows libraries to link their data with other rich datasets, creating a web of interconnected information. This enables users to discover related resources beyond their own library's holdings. For example: a library could link their catalogue data with other LOD repositories, to enhance search results. Searching for

"El Quijote" in the catalogue could return results not only from their own collection but also from other institutions that use LOD.

- **Increased Visibility:** By publishing data as LOD, institutions can increase their visibility on the web as researchers, developers, and other institutions can easily find and reuse library data. For example: Adding information about a rare copy of El Quijote in your collection to Wikidata would aid its discovery through Wikipedia articles (Libraries and Wikidata: Using linked data to expand access to library collections worldwide – Wiki Education).
- **Innovation**: LOD encourages creative applications and tools. Developers can build new services, visualisations, and applications using linked library data. For example: LOD allows the creation of new types of visualisations, such as timelines, maps and graph charts that can be useful to gain insight, in some cases without the need to install additional software thanks to the use of APIs. Some examples include:
  - a tutorial in Spanish to create map visualisations based on Wikidata and using several data repositories (e.g., members of the International GLAM Labs Community) as content
  - the exploration of machine-readable visual configurations to browse LOD repositories provided by Cultural Heritage institutions, including libraries, in the form of Jupyter Notebooks
  - a map representing the geographic locations mentioned in the metadata provided by a corpus of historical documents and paintings.

Though there are many benefits, SPARQL is the means by with Linked Open Data is queried and accessed and it's worth being aware that the use of API's based on SPARQL can be complex for less technical users since they need to understand how the data is modelled as well as be able to type a query. In addition, data quality has become crucial and several initiatives are focused on the assessment of the data quality provided by the catalogues.

Case Study: Manuscripts on Wikidata: the state of the art? | by Martin L Poulter | Medium

This example shows how to use Wikidata, a community-driven approach based on the Semantic Web and LOD that enables volunteers to edit the metadata, to describe manuscripts. It shows the expressivity of the vocabulary provided by Wikidata and the benefits of using Wikidata as a repository in terms of visibility and reuse.

The Jami' al-Tawarikh of Rashid al-Din [Hide]

object MSS 727 in the Khalili Collection of Islamic Art

Upload media

Instance of	manuscript
Owned by	Shahrukh (–1447)
Location of creation	Tabriz
Collection	Nasser D. Khalili Collection of Islamic Art (MSS 727)
Inventory number	MSS 727 (Nasser D. Khalili Collection of Islamic Art)
Inception	1314

Authority control [Hide]

Q107663668

Reasonator · PetScan · Scholia · Statistics · OpenStreetMap · Locator tool · Search depicted

Hands-on activity and other self-guided tutorial(s)

As part of my National Research Librarian's fellowship at National Library of Scotland exploring the adoption of Semantic Web technologies to transform, enrich and assess the Data Foundry's digital collections, I created a collection of Jupyter Notebooks that enables users to:

- understand the benefits of the adoption of the Semantic Web;
- create an RDF repository from a traditional dataset;
- enrich a dataset with external repositories such as Wikidata;
- reproduce the analysis and visualisations based on the datasets created.

I can also highly recommend starting with this Introduction to the Principles of Linked Open Data | Programming Historian tutorial which gives a great walk through of creating linked open data and includes an activity for using SPARQL to query LOD.

The course about Linked Open Data in cultural heritage collections, developed at Leiden University also includes a tutorial about a number of tools that can be used to create and to publish LOD. More specifically, it contains discussions of the LDwizard and CLARIAH Data Legend tool 'COW'.

To be able to retrieve and analyse Linked Open Data, you need to know how to build SPARQL queries. The following course can be helpful:

- Introduction to SPARQL

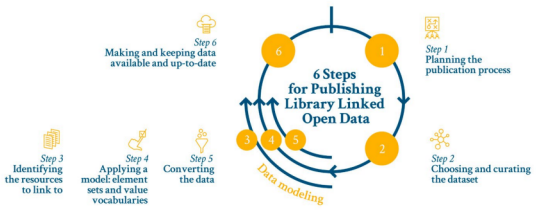
Examples of SPARQL queries used to collect and analyse data from heritage institutions can be found in the notebooks below:

- The Europeana SPARQL endpoint
- Wikidata
- Short Title Catalogue of the Netherlands
- The Dutch Institute for Art History

Recommended Reading/Viewing

If you are interested in learning more about LOD in terms of how to transform traditional bibliographic information into the Semantic Web, check out my research work performed as part of a fellowship at the National Library of Scotland in order to publish digital collections as LOD.

I can also recommend the excellent Best Practices for Library Linked Open Data (LOD) guide published by the LIBER Linked Open Data (LOD) Working Group in 2021 which outlines in detail six steps for publishing library linked data.



Overview of the Six Steps for Publishing Library Linked Open Data

Some examples of research articles to read providing additional details and information include:

- Towards a semantic approach in GLAM Labs: The case of the Data Foundry at the National Library of Scotland
- LIBER's Linked Open Data Working Group Publishes 'Best Practices for Library Linked Open Data (LOD) Publication' - LIBER Europe
- An automatic data quality approach to assess semantic data from cultural heritage institutions
- A Shape Expression approach for assessing the quality of Linked Open Data in libraries
- An Ontological Approach for Unlocking the Colonial Archive. (Example of transformation into RDF of a collection of maps using Open Refine.)
- Evaluating the quality of linked open data in digital libraries - Gustavo Candela, Pilar Escobar, Rafael C Carrasco, Manuel Marco-Such, 2022 (sagepub.com)

You can also find innovative ideas in the research articles published at the Semantic Web Journal.

Finding Communities of Practice

The LD4 Community is a community of practice for linked data in libraries.

Linked Art is a community working together to create a shared model based on LOD to describe cultural heritage with a particular focus on art.

Code4Lib is a community effort including a mailing list and a journal providing open articles based on the library domain and including LOD.