LINKED OPEN DATA:

Impressions & Challenges Among Europe's Research Libraries

Introduction

LIBER's Linked Open Data Working Group aims to paint a picture of the current state of Linked Open Data (LOD) among European research libraries and to provide insights which help research libraries to develop their LOD activities.[1]

The group recently completed a review of practices which research libraries follow in making data linked and open. The review was based on a survey which looked at processes for making data semantically interoperable.[2] Challenges and possibilities, from both a technical and librarian perspective, were also covered.

This document shares the survey results and explains how the results will be used in a future guide for libraries, offering linked data guidance and best practices.

Methodology & Respondent Profile

To avoid duplication, the survey questions were limited to topics that the group felt were not clearly or comprehensively addressed by other surveys. Results from other questionnaires will be used to add context when compiling best practices but the scope of this report is limited to the LOD Working Group survey.

Fourteen sets of answers were collected in spring-summer 2019. General information about each library's project was collected, as was information about tools, authority resources, linking, formats and schema (choices made, selection criteria and options considered but not used). Libraries were also asked to share helpful resources and to reflect on how LIBER could offer assistance for future LOD projects.

Of the 14 answers, approximately one third came from national libraries and two thirds from university libraries. Respondents were located in 10 countries: Canada, Estonia, Finland, France, Germany, Italy, Spain, Sweden, Switzerland and the United Kingdom.

Key Takeaways

The survey revealed that many libraries already use LOD in their processes. The following points were highlighted:

- **Linked data projects are diverse in their character and scope.** At the same time, there are certainly situations where it would have been possible to use less divergent approaches.
- **The most notable expense related to publishing linked data is human labour.** Providing guidance in the form of training and how-to guides is therefore of paramount importance.

1. https://libereurope.eu/strategy/research-infrastructures/linkedopendata
2. https://libereurope.eu/blog/2019/05/28/library-linked-open-data-survey

- **There is no one-size-fits-all tool.** A great variety of tools are used — commercial, open source and specialized — alongside locally developed routines.
- **The most commonly used vocabularies are GeoNames,[3] VIAF,[4] ISNI,[5] Wikidata,[6] and Dublin Core.[7]** Wikidata stood out as the most common external resource that the projects were linking to.
- **Data schemas used are often LOD-related: primarily SKOS[8] and Schema.org,[9] with mentions of FOAF[10] and Dublin Core as well.** A sizable minority opt for library-domain specific schemas like the Europeana Data Model[11] or BIBFRAME[12]
- **Libraries are keen to cooperate and exchange ideas.** This fits the character of LOD which intrinsically demands acting and thinking globally even when doing things at a local scale, however the networks enabling this are still somewhat thin on the ground.

Detailed Findings

The projects described were quite heterogeneous in scope, running the gamut from publishing all of the library's data in linked and open format to making a highly specific part of it available for a hackathon. The main goal for most was publishing library data in linked and open format (not surprising as it was the topic of the survey itself) but a couple goals were quite specialized (e.g., a map of the music scene or an ontology of emblems). The most common goal was publishing a sizable part (or even all) of the library's bibliographic data as LOD.

Roughly half of the projects were complete (6) and half were in progress (5) with three still in the planning stages. Most of the projects (10) published bibliographic data. Five also, or solely, published authority data (including vocabularies). Most of the LOD projects produced services and systems separate from the library systems but a sizable minority had managed to already integrate the LOD projects results into their own processes.

Depending on the nature of the project, its maturity and characteristics of the data (type, quality, amount), estimates differed regarding the workload needed for the publication project. Some projects included the publication of linked data as part of daily staff routines. Other projects had their own timelines and estimated person months separately.

The average workload for a publication project was 5-6 person months. It should be noted, however, that the size and scope of the projects varied greatly. The average should therefore be taken with a grain of salt, as obviously publishing a library's whole catalogue as linked data is very different from preparing for a hackathon.

Most projects reported that there were no other notable expenses beyond human labor. A small number of projects reported additional expenses including training costs, infrastructure, and outsourcing a third-party contractor for linked data training and/or mentoring.

3. https://www.geonames.org
4. http://viaf.org
5. http://id.loc.gov/vocabulary/identifiers/isni.html
6. https://www.wikidata.org/wiki/Wikidata:Main_Page
7. https://dublincore.org
8. https://www.w3.org/2004/02/skos
9. http://schema.org
10. http://xmlns.com/foaf/spec
11. https://pro.europeana.eu/resources/standardization-tools/edm-documentation
12. https://www.loc.gov/bibframe

Steps & Tools

The projects studied contained different steps such as data cleanup, processing and converting between various formats, data access, and so on: all of which require different tools.

Data cleanup was done mostly by using tools such as MarcEdit,[13] OpenRefine,[14] as well as more basic text and XML editors. SKOS is a widely adopted vocabulary standard and can be processed using various RDF tools and visualized using, for example, the SKOS Play! tool.[15] Some projects involved data using the MARC standard[16] and these were often processed using in-house routines and tools such as MARC Global,[17] and marc2bibframe2.[18]

Other tools used included: Metafacture[19] for data transformation, Elasticsearch[20] (a search and analytics engine for all types of data), Alma[21] (an integrated library solution), Catmandu[22] (a command line tool to access and convert data) and Virtuoso[23] (a 'data virtualization' tool). Overall, a great variety of tools are used — commercial, open source and specialized — alongside locally developed routines.

Authority Resources & Data Schemas

Depending on the task and project requirements, different authority resources were used. Some were selected for certain topics (places, languages, names). Others were used for more general subject indexing. The most commonly reported vocabularies were GeoNames, VIAF, ISNI, Wikidata, and Dublin Core. Wikidata stood out as the most common external resource linked to by projects. Linking is mostly done to Library and GLAM resources (VIAF, Dewey Classification, National Authority databases) and to general-interest resources like Encyclopedia Britannica. The LOD Cloud was mentioned as a possible source for finding new linking targets.

Data schemas were more often of the LOD variety, meaning SKOS and Schema.org with mentions of FOAF and Dublin Core as well. A sizable minority opted for more library domain specific schemas like the Europeana Data Model or BIBFRAME. The most common criteria for selecting formats were compliance with specific use cases or standards, making the data as usable and easy to access as possible. Other common criteria included openness and familiarity with the tools used.

Training & Guidance

All respondents said training or how-to guides for several aspects of linked data projects would be helpful. Specific attention was requested for the handling of bibliographic data by the different data models, handling of data and controlled vocabularies in multiple languages, and visualisation of linked data. Respondents also asked for an overview of 1) training resources 2) test cases and successful projects with details of

13. https://marcedit.reeset.net
14. https://openrefine.org
15. http://labs.sparna.fr/skos-play
16. https://en.wikipedia.org/wiki/MARC_standards
17. https://www.marcofquality.com/soft/mgfeatures.html
18. https://github.com/lcnetdev/marc2bibframe2
19. https://github.com/metafacture/metafacture-core
20. https://www.elastic.co
21. https://www.exlibrisgroup.com/products/alma-library-services-platform
22. https://librecat.org/Catmandu
23. https://virtuoso.openlinksw.com

their implementation 3) useful targets to link to and 4) tools and things to consider when selecting them.

The respondents shared many ideas in terms of how LIBER could help with their linked data endeavors. A common wish was for the creation of a governance mechanism to share the cost of linked data related work (e.g., alignments, code, etc). Needs were also expressed regarding the opportunity to cooperate through and participate in European funded projects and exchange ideas regarding linked data.

Lastly, many respondents expressed a need for training materials such as best practice documents, presentations of successful and unsuccessful projects, an update of the W3C LLD Incubator Group report, a directory of resources and library linked data workflows/ procedures.

Next Steps

These survey results will be compared and combined with results from other studies. From this, the Working Group will assemble best-practices and produce a basic workflow to guide institutions through the various steps of a LOD project. The group will also aim to include case studies highlighting how individual projects dealt with LOD and how that data was further used. Semantic interoperability is another topic which the group will aim to discuss in more depth.

The group aims to complete this second phase of its work by summer 2020 . If you would like to be involved, to be featured in a case study or to contribute in any way, please contact Matias Frosterus, Chair of the Working Group (*matias.frosterus@helsinki.fi*).