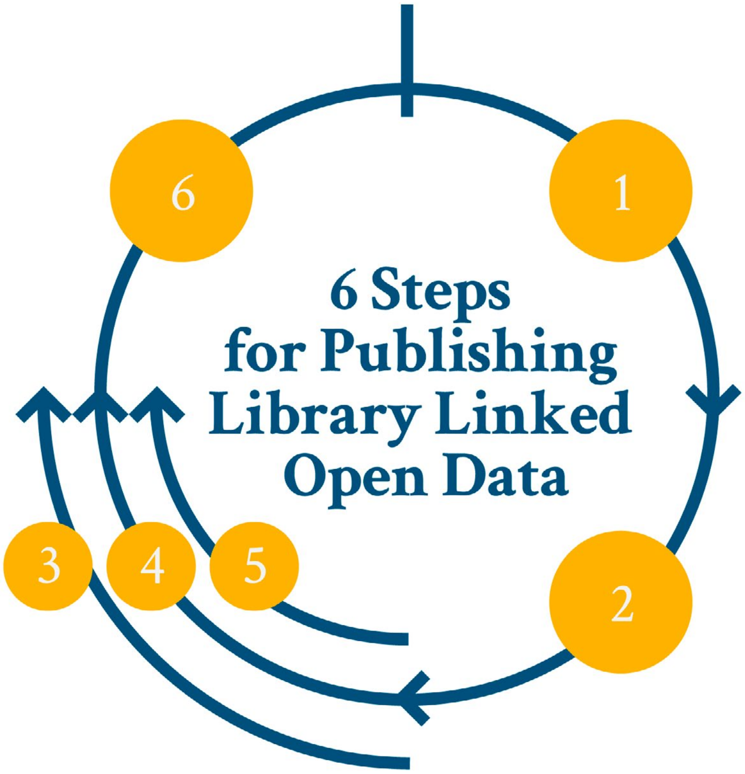


Best Practices for Library Linked Open Data (LOD) Publication

LIBER Linked Open Data (LOD) Working Group

February 2021 | www.libereurope.eu

Authors of this guide are as follows:

- Matias Frosterus (https://orcid.org/0000-0002-8355-0256)
- David Hansson (https://orcid.org/0000-0003-0835-4367)
- Maral Dadvar (https://orcid.org/0000-0002-1351-2561)
- Ilias Kyriazis (https://orcid.org/0000-0001-8958-0168)
- Sofia Zapounidou (https://orcid.org/0000-0001-8784-9581)
- Friedel Grant

Read more about LIBER's Linked Open Data Working Group.

About LIBER

LIBER (Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries) is the main network for research libraries in Europe. Founded in 1971, LIBER has grown steadily to include more than 400 national, university and other libraries from over 40 countries.

Together we work to represent the interests of European research libraries, their universities and their researchers by advocating on issues such as Copyright and Open Access, by collaborating on European-funded projects, and by meeting and learning at events such as our Annual Conference.

www.libereurope.eu

Contents

1 Introduction

Linked Open Data (LOD) is becoming an increasingly popular way of publishing data for others to use. The principle behind it is very simple. Firstly, there are different types of information and concepts, as well as the relationships between them. In this context, we will refer to these three components as 'things'. When you publish data we, 1) use identifiers for all these 'things', 2) make the identifiers functional so they can be used to access these 'things'. And, 3) provide useful metadata about the 'things' when you access them through their identifiers. As the term 'linked' implies, the 'things' from multiple resources can be linked to each otherresulting in an interconnected web of information which can be easily machine-processable. When the result is published for external and free use (using an open license), the result is LOD[1].

This document looks at publishing LOD from a library perspective and argues why it should be employed and how. We will not delve into technical details nor their respective technical tools. Instead, we will present various aspects of the topic, introduce different options available, and lay out a foundation for possible exploration at a later stage. As such, the body of this document presents the six steps of LOD publication and explains each one of these steps in depth. Thereafter, suggested readings are provided to help you delve deeper into the subject, if needed.

2 LIBER Linked Open Data Working Group

LIBER's Linked Open Data Working Group was founded with the goal of examining state-of-the-art practices in the field of lLOD publishing and its respective development within research libraries. This document is therefore the result of this group's work from 2018 to 2020. Additionally, the group has held workshops on LOD at various LIBER Annual Conferences.

When it comes to LOD, LIBER acts as a network for collecting experiences and facilitating discussions. The articles or reports written under the LIBER umbrella target a wide audience. This means that our shared considerations as well as possible solutions can potentially enhance interoperability and make the re-use and cross-use of data tangible and implementable. This, in turn, makes library data more attractive to both end-users and developers outside of the library sector.

3 Survey on Linked Open Data Publication for Libraries

A starting point for this document was a survey carried out by the LOD Working Group. This survey studied current LOD activities carried out by European research libraries. The results of the survey made it clear that many libraries already use LOD. Below, you will find the key points brought about by these survey results:

[1] For a more precise and robust definition, this report adopts the definitions of library linked data as provided by W3C LLD Incubator Group reports: https://www.w3.org/2005/Incubator/lld/

- Linked data projects are diverse in character and scope;
- The most notable expense related to publishing linked data is human labour;
- There is no one-size-fits-all tool. A great variety of tools — commercial, open source and specialized ones — are used alongside locally-developed tools;
- The most commonly used vocabularies are GeoNames, VIAF, ISNI, and Wikidata;
- The used data schemas are often LOD-related: primarily SKOS and Schema.org, with FOAF and Dublin Core also mentioned;
- Libraries are keen to cooperate and exchange ideas.

4 Why Should Libraries Publish their Data in an Open and Linked Format?

The main benefit of publishing library data as LOD is that it makes data readily available and easier to use for researchers, system developers, librarians etc. The LOD format makes data more attractive and is easier to analyze, combine, and integrate.

Moreover, making the data 'linkable' allows a user to enrich it with the help of external resources. This enrichment can be accomplished by adding, for example, missing data (e.g. the missing year of death of a person) or new information (e.g. geographic coordinates) to the data pool. Links can also be used for finding discrepancies in the data and thus correcting mistakes.

An obvious challenge when determining the success of an LOD project is that, once your data is open, it can be difficult to track who is using it. ou can, of course, track downloads, API use, and other web statistics, but these will not give you much insight into how the data is used. You can also encourage users to tell you what they are doing with your data, and such word-of-mouth reports may in fact comprise most of the insights you are able to gather. Searching for links to your identifiers can also show who is using your data, but measuring success still remains a challenge.

Below you will find an example showing the use and benefits of LOD for libraries:

5 Example Project: CERL's Resources as Linked Open Data

The Consortium of European Research Libraries (CERL) hosts a number of databases on Early Modern book history. It also maintains the CERL Thesaurus (CT), a database that serves as a central resource, both for use in CERL's own databases and beyond. This database collects corporate and personal names, as well as associated names of printing places and printers, primarily from imprints of Early Modern books. The CT, as of June 2020, holds 1.383.482 records and is frequently updated.

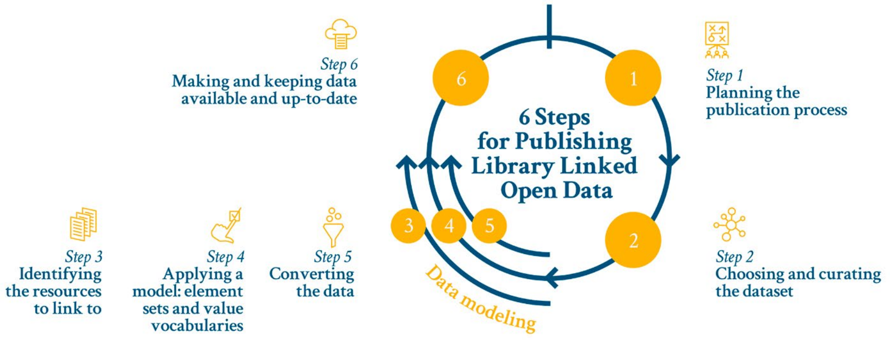
Publishing the CT as LOD is not a single, one-off project. Rather, LOD is seen as a continuous process of providing access to data in accordance with the latest standards. Note that the CT was established in 1999 (with the technological limitations of that time) but has always been aligned to values of openness and the connectedness of LOD. In recent years, the CT has been linked to several other data sources such as the German National Library's Gemeinsame Normdatei (GND) and Wikidata.

Beyond the day-to-day use of databases accessible through the bibliographic research community's own interfaces, CERL's datasets are also being used within a variety of Digital Humanities projects. The most prominent of these is the CERL Thesaurus which is used by multiple projects in order to normalize, for example,the recorded names of printersames. Aside from the CT, the Heritage of the Printed Book database has become a focus of attention for bibliographic data science and metadata quality analysis. Its prominent use also indicates where such efforts will be focused in the future.

For a more detailed description of the above-mentioned CERL project, including a discussion on challenges and recommendations based on the collected experience, see the Appendix.

6 Six Steps for Publishing Library Linked Open Data

The LIBER LOD Working Group has developed six steps when it comes to publishing library LOD. For each one of these steps we share related practices and decision points, potential actions to take, and useful hints. This figure below gives an overview of the steps and thereafter we outline each step in further detail:


Step 6
Making and keeping data available and up-to-date
6
1
Step 1
Planning the publication process
6 Steps for Publishing Library Linked Open Data
3
4
5
2
Step 3
Identifying the resources to link to
Step 4
Applying a model: element sets and value vocabularies
Step 5
Converting the data
Step 2
Choosing and curating the dataset
Data modeling

Overview of the Six Steps for Publishing Library Linked Open Data

It is worth noting that the depicted progression of the above steps is not strictly linear. Iterations to revisit steps can occur, and new adjustments can be made. This aspect specifically applies to the three steps of the 'data modeling' process, which are highly intertwined and may run simultaneously rather than sequentially. Moreover, it should be noted that not all of the six steps are always required. If, for example, you are building a LOD hub from scratch, there is obviously no previous data that requires conversion. Hence, the steps can serve as a guideline which can be adapted accordingly to LOD needs.

Step 1 - Planning the Publication Process

Planning the project is always the first thing to do when getting started. Because linked data can be complex for many, starting small is the best way to gain experience and feel comfortable with the various publication stages. The following points need to be taken into account at this stage:

1. Scope of the project: Defining the aim of the project is critical. Do you want to acquire experience? Gain interoperability with other datasets? Or, are you simply testing the waters, exposing your own dataset to see if anyone would use it for their own purposes?

2. People and expertise: The project team needs to have the knowledge to handle several tasks. These include[2]: Choosing a dataset, modeling the data, describing the data with standard vocabularies, deciding how to present the data using URIs, converting the data, providing machine access to data, choosing a publication license, announcing the published datasets, taking care of stakeholder communication, and recognizing the social contract (that is, keeping the data available and updated once published). Interpersonal skills to communicate with team members, as well as a good knowledge of the dataset(s) at stake are also important. Knowing how to model the data requires, first and foremost, knowledge of the domain the data represents. Using multiple vocabularies is also a significant skill, which requires a broader understanding of the information ecosystem, as well as domain expertise. If the selected dataset demands extended curation, the team member with the most relevant skill set should be easily available.

3. Tools: The team should be familiar with the relevant tools for converting the data. There can be several possibilities: from programming your own scripts to using open source or commercial software. The same applies when cleaning up your dataset(s).

4. Resources: Workload and timing of the project should be well planned. According to the survey “Linked Open Data: Impressions & Challenges Among Europe’s Research Libraries”[3], which was pursued by this group, the respondents claimed that labour costs made up the major part of the resources required in their linked data projects.

5. Steps and milestones: Setting steps and defining milestones will help you get through the project more easily and in a structured manner. Apart from the steps described in this document, smaller accomplishments such as choosing and finding the right tools, forming the project team, and setting a workflow, will be important during the project.

[2] Based on the W3C Best Practices document for Publishing Linked Data: https://www.w3.org/TR/ld-bp/
[3] https://zenodo.org/record/3647844

Step 2 - Choosing and Curating the Dataset

Even with the scope of the project having already been clarified, there are still considerations when choosing the dataset(s) for linked data transformation and publishing. According to the W3C LLD Incubator Group report[4], the term 'dataset' refers to a set of 'library-related resources'. As an example, a dataset may include bibliographic data extracted from catalogs (data from bibliographic records or authority files), a local value vocabulary or data related to a specific collection. The current section describes the criteria of dataset selection, as well as the necessary steps for its curation and clean-up.

Dataset selection

In addition to the factors influencing the development of the project in general (experience, expertise, resources, goals, etc.), the following points need consideration:

- Content: What does the dataset include? Is it authority or bibliographic data? Is it a digital collection?
- Systems: From which system or systems (e.g., online catalogue, repository, database, API interface, etc.) is the data going to be extracted?
- Types of entities used: What types of entities are being used? Are they referring to person/organization names (authority data), are they bibliographic (bibliographic data), or are they describing digital objects?
- Metadata schemas: What type of legacy metadata (e.g. MARC21, Dublin Core, MODS, VRA, EAD, TEI, etc.) is used?
- Size: How big is the dataset?
- Data quality: Is the dataset reliable and complete? Is it consistent? Are inconsistencies random or systematic?
- Uniqueness: How unique is the dataset? The more unique, the more added value it will have, when linked to other datasets in the linked data cloud.
- Popularity: Is the dataset popular to your users? Will its publication as linked data add additional value and enable further research? Will you consider feedback from users regarding the selection of the dataset or its future uses?
- Reuse: Is there a high reuse potential? The linked Name Authority file of a National Library, for example, will have - most probably - more possibilities of being used compared to a dataset of narrower scope.
- Ownership: Who owns the data? In case of MARC21 authority and bibliographic records, many of them have been imported from other libraries or vendors. Make sure the related rights have been clarified and cleared, and there are no restrictions that could hinder the publication of the dataset as LOD.
- Privacy: Are there any privacy issues? Datasets with privacy issues must be anonymized or excluded from the linked data publication process.

[4] The appendix of the W3C LLD Incubator Group Final report provides three categories of library linked data: datasets, value vocabularies, and metadata element sets. https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/#Appendix_A:_An_inventory_of_existing_library_Linked_Data_resources

Data curation and clean-up

Based on the inspection of the dataset and the identification of user expectations or possible quality gaps, the next step is to curate the dataset and clean it up for further adjustments and subsequent transformations. Cleaning up may require significant resources, thus the selection of the dataset may depend on the extent of curation and the number of employees available to perform the task.

Data cleanup involves fixing errors and inconsistencies, removing whitespaces, correcting differing formats (e.g. of dates), removing duplicate information, adding lacking data, etc. It must be noted that it is a matter of policy if the data cleanup will be done to the original data or to the data that will be eventually published as linked data. Nevertheless, every step of the cleanup process must be well-documented to enable the implementation of the data cleanup workflow to other datasets or the improvement of the data cleanup process in case problems occur. The tools to be used can be either open source or proprietary. The following list mentions some open-source ones:

- MarcEdit: a MARC editing utility developed by Terry Reese. Used for cleaning up MARC records.
- Spreadsheet: a set of 'School of Data' guidelines to cleanup datasets using spreadsheets.
- OpenRefine: it may be used for cleaning data in tabular form (e.g., xls, csv) or in XML.
- Datacleaner Community Edition: it may be used for relational databases and data in tabular form (e.g., xls, csv).

Note: Even though the cleaning up of data takes place as preparation for the modeling process, decisions can always be revisited, depending on the later characteristics and constructs of the selected model.

Step 3 - Identifying the Resources to Link to

The Semantic Web aims to create links between data and make these links understandable by machines. As mentioned previously, LOD refers to a set of principles that make these links possible and afford accessibility of the interlinked data on the Web. The more entities (e.g., things, events, people, locations) are connected together, the more powerful, comprehensive and expandable the data can be.

However, to link and integrate datasets from different resources depends very much on the project goals and the topic of interest. Selecting the right external source depends also on the information you need to add to your dataset. So, if, for example, you want to add geographical coordinates, you need to link the dataset to sources that could offer such information, like GeoNames[5], which provides LOD-format data and can offer semantic value via its ontology.

[5] http://www.geonames.org/ontology/documentation

Apart from the target linking between local data and external ones, there are also specific procedures, where linking is used for data refinement purposes. Reconciliation, for example, involves the replacement of local values to their respective ones, as taken from well-known controlled vocabularies, like the Library of Congress Subject Headings (LCSH), or the Art and Architecture Thesaurus (AAT). Further enrichment is also accomplished by the recognition of named entities (Named Entity Recognition) and the addition of their URIs to the dataset. Both processes may be executed with OpenRefine extensions. should also be noted that the number of open source tools, especially in the case of NER, is rather limited.

Frequently used datasets by libraries

NAME OF THE DATASET	DESCRIPTION	SOURCE
American Numismatic Society's nomisma	A thesaurus of numismatic concepts. he 2015 survey, reported daily usage between 10,000 and 50,000 requests a day. So usage has more than doubled over the last three years.	http://nomisma.org/
Bibliothèque nationale de France	It provides access to the BnF collections and provides a hub amongst different resources. In the 2015 survey it reported daily usage of between 10,000 and 50,000 requests a day.	data.bnf.fr
Europeana	Aggregates metadata for digital objects from museums, archives, and audiovisual archives across Europe. It reported the same daily usage in 2015.	europeana.eu
Library of Congress	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
National Diet Library	National Diet Library's NDL Search, providing access to bibliographic data from Japanese libraries, archives, museums and academic research institutions. It reported the same daily usage in 2015.	iss.ndl.go.jp
North Rhine-Westphalian	North Rhine-Westphalian Library Service Center's LOD, provides access to bibliographic resources, libraries and related organizations, as well as authority data. It reported the same daily usage in 2015.	lobid.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. It reported the same daily usage in 2015.	viaf.org
WorldCat	OCLC's WorldCat Linked Data, a catalogue of over 400 million bibliographic records made experimentally available in linked data form. It reported the same daily usage in 2015.	Worldcat.org

Table 1: OCLC survey: used datasets by libraries

According to the ‘International Linked Data Survey for Implementers’ conducted by OCLC[6], the eight most heavily-used linked data datasets by libraries (as measured by the average number of requests per day) are presented in Table 1. These resources were also studied in an earlier survey by OCLC[7] in 2015.

Additionally, another survey conducted by the Association of European Research Libraries (LIBER[8]) highlights the following resources (Table 2) as the most commonly used ones.

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

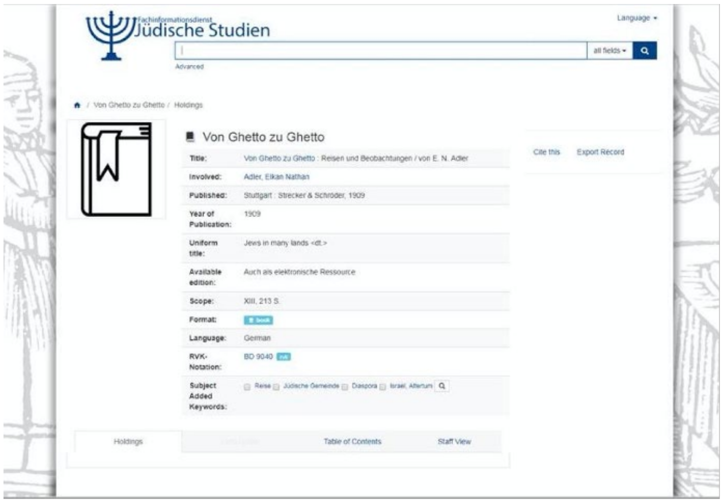
Table 2: LIBER survey: Frequently used datasets by libraries

Certainly, there are many more resources to link to, depending on the purpose and scope of your project. Some of these resources cover a wide range of topics, while others are centered on a specific subject, for example Judaicalink[9], which is a specialised knowledge base for Jewish studies.

[6] https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-survey
[7] http://www.dlib.org/dlib/july16/smith-yoshimura/07smith-yoshimura
[8] https://libereurope.eu/blog/2020/02/06/linked-open-data-impressions-challenges-among-europes-research-libraries
[9] http://www.judaicalink.org

A linking example

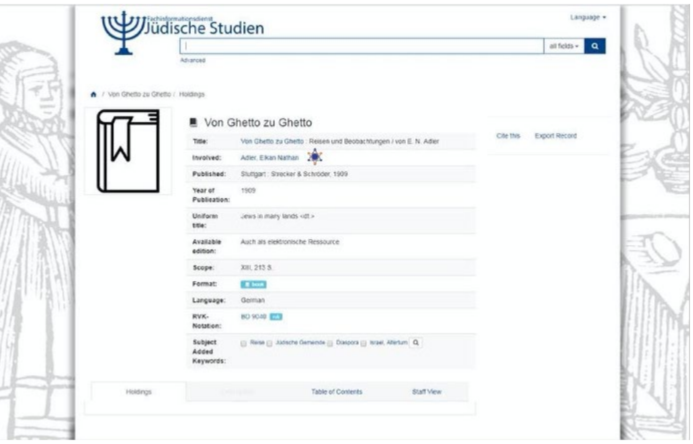
To clarify how external resources can provide further information when linked, we will look at an example from the Goethe University Library (Frankfurt, Germany) project titled, FID Judaica[10]. Part of this project is to gain additional information about authors of the library's Judaica collection and provide this information to users when searching for a certain title. For this purpose, Judaicalink has been used. Prior to using Judaicalink, when a title was searched, the triggered result would have displayed only the author's name. However, afterwards, the collection's authors have also been enriched with additional information. Hence, the search results also display the Judaicalink logo as well. By clicking on it, extensive information about the author is now visible:



Von Ghetto zu Ghetto

Title:	Von Ghetto zu Ghetto : Reisen und Beobachtungen / von E. N. Adler
Involved:	Adler, Elkan Nathan
Published:	Stuttgart : Strecker & Schröder, 1909
Year of Publication:	1909
Uniform title:	Jews in many lands <dt.>
Available edition:	Auch als elektronische Ressource
Scope:	XIII, 213 S.
Format:	Book
Language:	German
RVK-Notation:	BD 9040
Subject Added Keywords:	□ Reise □ Jüdische Gemeinde □ Diaspora □ Israel, Altertum

Figure 2a (above): The search results prior to linking to Judaicalink.

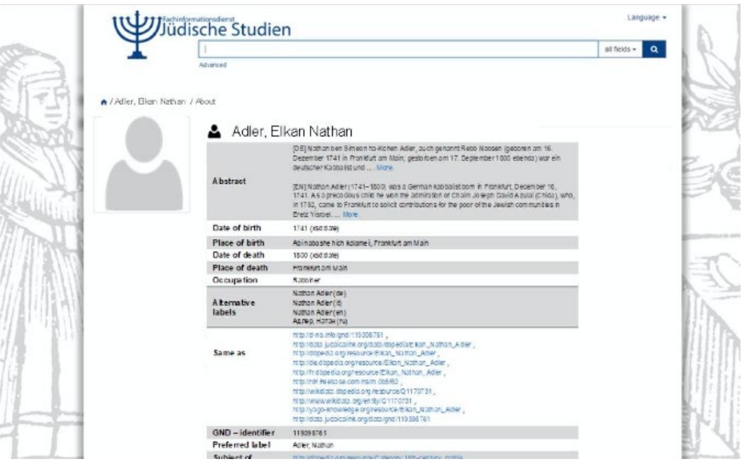


Von Ghetto zu Ghetto

Title:	Von Ghetto zu Ghetto : Reisen und Beobachtungen / von E. N. Adler
Involved:	Adler, Elkan Nathan
Published:	Stuttgart : Strecker & Schröder, 1909
Year of Publication:	1909
Uniform title:	Jews in many lands <dt.>
Available edition:	Auch als elektronische Ressource
Scope:	XIII, 213 S.
Format:	Book
Language:	German
RVK-Notation:	BD 9040
Subject Added Keywords:	□ Reise □ Jüdische Gemeinde □ Diaspora □ Israel, Altertum

Figures 2b and c (above and on the page that follows): The search results after linking to Judaicalink and enriching authors with additional information.

[10] https://www.ub.uni-frankfurt.de/judaica/home_en



Fachinformationsdienst Jüdische Studien

Language ▾

all fields ▾

Advanced

/ Adler, Elkan Nathan / About

Adler, Elkan Nathan

Abstract	[DE] Nathan ben Simeon ha-Kohen Adler, auch genannt Rebb Noosen (geboren am 16. Dezember 1741 in Frankfurt am Main; gestorben am 17. September 1800 ebenda) war ein deutscher Kabbalist und ... More [EN] Nathan Adler (1741–1800) was a German kabbalist born in Frankfurt, December 16, 1741. As a precocious child he won the admiration of Chaim Joseph David Azulai (Chida), who, in 1752, came to Frankfurt to solicit contributions for the poor of the Jewish communities in Eretz Yisrael. ... More
Date of birth	1741 (xsd:date)
Place of birth	Abinaboshe hich Kolomei, Frankfurt am Main
Date of death	1800 (xsd:date)
Place of death	Frankfurt am Main
Occupation	Rabbiner
Alternative labels	Nathan Adler (de) Nathan Adler (it) Nathan Adler (en) Адлер, Натан (ru)
Same as	http://d-nb.info/gnd/119398761 , http://data.judaicalink.org/data/dbpedia/Elkan_Nathan_Adler , http://dbpedia.org/resource/Elkan_Nathan_Adler , http://de.dbpedia.org/resource/Elkan_Nathan_Adler , http://fr.dbpedia.org/resource/Elkan_Nathan_Adler , http://rdf.freebase.com/ns/m.0b5452 , http://wikidata.dbpedia.org/resource/Q1170731 , http://www.wikidata.org/entity/Q1170731 , http://yago-knowledge.org/resource/Elkan_Nathan_Adler , http://data.judaicalink.org/data/gnd/119398761
GND – identifier	119398761
Preferred label	Adler, Nathan
Subject of	http://dbpedia.org/resource/Category:18th-century_rabbis

Step 4 - Applying a Model: Element Sets and Value Vocabularies

One of the reasons for publishing library data as linked data is to free data out of MARC (or equivalent) silos. On the other hand, the danger of creating LOD silos is also present[11], and this is a good reason why libraries need widely-known models for describing their resources (wherever possible).

Even if there is no single model that fits it all, selecting one is strongly recommended. In bibliographic data such models depict bibliographic entities and their attributes, as well as their relationships to other entities. These models may be described in documents, or may be formally expressed in languages such as RDF Schema or OWL. Within this context, bibliographic entities are modeled and defined as classes, whereas attributes and relationships are modeled and defined as properties.

By applying models to data publication, the bibliographic entities, their attributes and their relationships will be better understood both in the context of your library and outside of it, since there will be external definitions and references. Such widely understandable data will therefore present a higher reuse potential.

Bibliographic models: IFLA LRM, RDA and BIBFRAME 2.0

The IFLA Library Reference Model (former FRBR) is developed by IFLA and consolidates the FRBR family of models, namely FRBR, FRAD, and FRSAD. The official metadata element set has been published[12]. The former FRBR model was expressed in RDF and remains available in the Open Metadata Registry.

The RDA (Resource Description and Access) model is maintained by the RDA Steering Committee. It adheres to the principles and conceptualizations of the IFLA LRM model, but it also provides refinements to many IFLA LRM attributes and relationships. RDA element sets are available through the RDA Registry.

[11] https://www.o-bib.de/article/view/2017H2S1-13
[12] https://www.iflastandards.info

BIBFRAME 2.0 is maintained by the Library of Congress. It takes into account IFLA LRM conceptualizations, but presents significant modeling differences to both LRM and RDA. The BIBFRAME ontology is available as an RDF file through the BIBFRAME website.

Use cases: The FRBR model has been used by the National Libraries of Spain, France, and Iran in their linked data projects. The RDA content standard is the de facto standard in libraries worldwide used to include FRBR semantics in MARC records. RDA is also used to prepare MARC records for future conversions to linked data. Its element sets have been mostly used in FRBR implementations. The BIBFRAME model seems to be gaining popularity, probably thanks to the fact that it's a Library of Congress model and there is an active community supporting it. BIBFRAME has been tested by the National Libraries of Sweden, Finland, Germany, and by the COBIS initiative in Italy.

Element sets

Element sets provide the elements for describing a model's primitives. They can be model-specific (e.g. the RDA Element Set defining the RDA classes), while others may be generic and therefore used in diverse model instances (e.g. the Dublin Core Metadata Element Set). Selecting well-known metadata element sets will likely make your dataset more accessible.

The following is a list of element sets being used mostly in library linked data projects:

USE	PREFIX	COMMENTS
GENERIC	rdfs	Rdf Schema
	dcterms	Dublin Core
	schema.org	
	cc	Licensing terms
BIBLIOGRAPHIC DATA	isbd	ISBD
	bibo	Bibliographic ontology
	rda	Resource Description & Access
	bf	BIBFRAME
AUTHORITY DATA	foaf	Description of people
	skos	Description of concepts
	org	Description of organizations
	bio	Biographical information
	mads	MARC authorities
	gnd	DNB vocabulary for authorities

Table 3: Frequently used element sets in library linked data projects, according to the OCLC 2018 International Linked Data Survey for Implementers, the LIBER Linked Open Data Working Group survey, and the Kim Tallerås' study on the quality of linked bibliographic data

To discover metadata element sets (MES) you will need to consult lists, specialized search engines, and metadata registries. Here are some suggested resources:

- Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets #Metadata Element Sets: This list was published in 2011 and has not been updated ever since.
- Linked Open Vocabularies (LOV): This registry is regularly updated.
- Open Metadata Registry: In this registry you can search for metadata element sets and single elements to suit your needs. Regularity of updates depends on the contributors. You may need to further search for the current version of a metadata element set, schema, or value vocabulary at their projects' webpages.
- prefix.cc: namespace lookup for RDF developers: Helpful resource if you find a prefix in a dataset and do not know what it means.
- Linking Open Data Community > Common Vocabularies: A list of common vocabularies by the W3C Linking Open Data Community.

To evaluate if a MES serves your dataset's descriptive needs, first consider the following points, as suggested by the Linked Data handbook:

- Usage: Is the MES used by other library-related linked data projects?
- Maintenance: Is the MES regularly maintained and updated?
- Coverage: If the MES is model-bound, does it cover the selected dataset's descriptive needs? If the MES is a generic one, can it cover the dataset adequately?
- Expressivity: Is the MES too expressive or too generic? Is the MES's expressivity adequate to describe your dataset?

Value vocabularies

In a linked data environment, consistent and structured information is needed. Consider providing values to elements using controlled vocabularies. This adds coherency to your dataset and provides more options for possible linking to external resources.

Note that there are many value vocabularies, varying in scope, detail and purpose. The following table provides a list of well-known ones. Similarly to the models and the metadata element sets, the selection of value vocabularies depends on the data and the scope of your project. Keep in mind that the use of commonly-used value vocabularies will most possibly increase the understandability and usability of your dataset.

USE	NAME
Classification	UDC
	LCC
Subject / Thesauri	LC Subject Headings
	MeSH
	RAMEAU
	Agrovoc
	Art & Architecture Thesaurus (AAT)
	Eurovoc
	Getty Thesaurus of Geographic Names
Name authority data	ISNI
	VIAF
	LC Name Authority File
	ULAN
	Geonames
Other	MARC Code List for Relators Scheme
	ISO 639-2 Languages
	Creative Commons

Table 4: Well-known value vocabularies

Step 5 - Converting the Data

Once you have your data model and you have cleaned up your data, the next step is to convert the data to the chosen schema and make this available in a format suitable for linked data. Usually, the original source data is kept and the conversion creates a separate set of linked data. The specific combination of original data and target schema is often somewhat unique, so it might be difficult to find an off-the-shelf tool that will do exactly the conversion you need. However, there are many open source solutions that, with a bit of configuration work, will either outright perform the needed transformations or work as a starting point for further refinements. In general, the choice of a specific tool should be guided by whomever is going to do the actual work and the languages they will be most comfortable with, be it Python, XSLT, Java, or another language. The end result is some serialization of RDF with JSON-LD and Turtle being the most popular currently.

No matter the exact implementation chosen for the conversion, you'll need a mapping from the original format to the new one. Based on the output of data modeling, you should create a simple spreadsheet detailing the matching properties and elements in the original format versus the ones in the new format. It is possible that this mapping will not be completely straightforward and that you will end up with a moderately complex set of rules, e.g. "if this MARC subfield has this value, then the value from this field is a modifier for the value from that field."

If you have worked on such a conversion before and you are using a ready-made tool, this step will be a lot simpler. However, even in this case, you should review the conversion logic and make sure it fits your current data.

Aside from matching and converting the properties, you might also need to map and convert the values. The original data might have, for example,persons, places, and subject headings that are referred to using a unique naming scheme. For linked data, these naturally need to be converted so that the values are identifiers that refer to the dataset about persons, places, or subject headings. This mapping is not usually feasible to be done by hand onto a spreadsheet but rather you should have the conversion programme matching the data.

Identifiers: One important consideration in the conversion is the planning and implementation of identifiers for all the things that are being published as linked data. As noted before, the identifiers should be functional: they should grant access to metadata about the things themselves. They should also be stable. The most light-weight approach are Cool URIs[13] - stable URIs that are formed in such a way that they do not need to change even if something about the object they refer to changes. A more robust approach is to use proper PIDs (persistent identifiers) such as URN or Handle. For more information on the various PID systems and their use in the cultural heritage domain, see Lukas Koster's article Persistent identifiers for heritage objects[14].

Once the conversion is done, the result should be verified in some way. A manual inspection of the result is the first step - selecting some samples from the data and making sure that the result is what it was meant to be. Here it would be very useful to employ the original experts on the data, to have them make sure that the data is as it should be.

Step 6 - Making and Keeping Data Available and Up-Toto Date

Once you have your LOD dataset ready, the final step is to publish it. Three concerns are prominent here: how do you serve your data, how do you license this, and how do you make everyone aware of the existence of the data?

How to serve your data

Serving your data depends on the amount of data you have, but offering a simple downloadable bulk file is usually a good practice. It is very easy to set up and quite attractive for users who want to access the data. If the data is available in several formats, consider which ones could be of good use and therefore should be published.

Having a SPARQL[15] endpoint for querying the data is also a great idea. SPARQL allows the user to extract information via querying and in effect provides a "programmable" API to the data.

[13] https://www.w3.org/TR/cooluris/
[14] https://journal.code4lib.org/articles/14978
[15] https://www.w3.org/TR/sparql11-overview/

With a traditional API the user can only access the data in the ways provided by the API, whereas with a SPARQL endpoint users can query the data in any way they need: SPARQL allows you to traverse the links between resources efficiently, and stringing together queries with various functions allows delving into the data in a sophisticated way. However, due to the language being so powerful, it is quite easy to make very intensive queries, and for this reason some care should be taken. A SPARQL endpoint requires a little more work to set up than a simple dump and additionally presupposes a triplestore[16] of some sort. Many ready solutions exist for this, and putting the data into a triplestore should not be very difficult.

Documentation and license

Aside from the data itself, you should also include documentation about it. This can be done somewhat formally and in a machine-processable way by using the Vocabulary of Interlinked Datasets (VoID)[17] and simply providing a human-readable description of the dataset including information produced within the previous steps should be enough.

One important aspect when making data available is also deciding on licensing. LOD as a term implies the use of an open license, and there are some specific options one can choose from. The simplest one is to go for a Creative Commons (CC) license, which makes the use of data simpler as opposed to having a customized open license. CC0[18] is the most permissible of the CC licenses but others can be used as well. It is also good practice to include the license information in the content itself.

Making your data known

After the publication is done, let the world know about your data! Work with your library’s communication team to get the word out. Tap into communities on Twitter, and offer your dataset to hackathons and similar events.

It is becoming also common practice for libraries to produce their own data catalogues - a website that lists the datasets and APIs the library offers, as well as some guidance on how to use them. When setting up a data catalogue, consider utilizing DCAT[19], a vocabulary for describing datasets. Apart from setting up your own data catalogue, there are also general data hubs and repositories for various types of LOD data. For example, if you are publishing a vocabulary, Bartoc lists resources from all over the world; making sure that your data is included in it also increases its visibility.

Maintaining the data

Finally, it is important to note that the work does not end at publication. The data needs maintenance and it is common to publish an updated version every so often depending on the nature of the data.

[16] A specific type of database meant for RDF data.
[17] https://www.w3.org/TR/void/
[18] https://creativecommons.org/share-your-work/public-domain/cc0/
[19] https://www.w3.org/TR/vocab-dcat/

This means that when performing the publication steps, their repeatability should also be considered. For example, if there has been manual correction steps, the results from those should be preserved in a way that can be automatically repeated when a new version of the data is to be published. Even if the contents of the data remain stable, it is still important to periodically check that all the links to external resources are still working. This means that it is usually a good idea to have a maintenance plan and schedule for each LOD dataset which makes sure that the data is kept up to date and high quality.

7. Other resources

In close, by now it is clear that we have only just touched upon this important topic for libraries. There is a wealth of excellent information out there to help you delve deeper into the topic. One great starting point is LD4L (Linked Data for Libraries) which demonstrates several years of work collecting and producing resources and tools to help with LOD publishing for libraries. Another great resource is library conferences which offer a chance to exchange ideas and learn from others' experiences. The main conferences for library linked data concerns are SWIB[20] and ELAG[21] but many other conferences also feature topics related to library LOD.

Outside of the library world, the Linked Open Data Cloud is also certainly worth mentioning. It collects a huge number of interlinked datasets and is a well-known place for submitting your dataset and gaining visibility. The cloud was somewhat dormant for a while, but lately it has been picking up again.

Finally, we are including a more targeted list of suggested reading for further guidance. The list is by no means comprehensive. Each of the entries is, however, worth reading. The list is presented in no particular order and you can find it at the very end of this document.

8. Conclusion

It should be said that everything depends on the data and the scope of the project. If the project is small, or if it is following in the footsteps of someone who has done something very similar, some of the steps might be almost trivial. On the other hand, if there are many different types of data or if the ambition of the project is high, some of the steps might be quite laborious. So, proper planning goes a long way and considering the steps we have outlined here will hopefully make your planning a little simpler.

[20] https://swib.org/
[21] https://elag.org/

9. References

- 7 Data Cleanup Terms Explained in Plain English - Rapid Insight Inc. -. (2020). Retrieved 15 May 2020, from https://www.rapidinsight.com/blog/7-data-cleanup-terms-explained-plain-english/

- Alemu, G., & Stevens, B. (2015). An Emergent Theory of Digital Library Metadata: Enrich then Filter. Chandos Publishing.

- Baker, T., et.al. (2011).Library Linked Data Incubator Group Final Report: W3C Incubator Group Report 25 October 2011. https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

- IFLA Study Group on the Functional Requirements for Bibliographic Records. (2009). Functional Requirements for Bibliographic Records Final Report. Retrieved from http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

- Isaac, A., Waites, W., Young, J., & Zeng, M. (2011). Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets. Retrieved from https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

- Koster, Lukas. "Persistent identifiers for heritage objects." Code4Lib Journal 47 (2020). https://journal.code4lib.org/articles/14978

- Library of Congress. (2016). Overview of the BIBFRAME 2.0 Model. Retrieved April 14, 2018, from https://www.loc.gov/bibframe/docs/bibframe2-model.html

- Riva, P., Bœuf, P. Le, & Žumer, M. (2017). IFLA Library Reference Model: A Conceptual Model for Bibliographic Information. Den Haag. Retrieved from https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf

- Taylor, H. (2011). "Ownership" of MARC-21 records. Retrieved 15 May 2020, from http://cul-comet.blogspot.com/p/ownership-of-marc-21-records.html

10. Further Reading

- A Beginner's Guide to Creating Library Linked Data: Lessons from NCSU's Organization Name Linked Data Project Hanson, Eric M. Serials Review, 10/02/2014, Vol.40(4), pp.251-258"

- Bauer, F., & Kaltenböck, M. (2012). Linked open data: the essentials. Wien: Ed. mono/monochrom.https://www.reeep.org/LOD-the-Essentials.pdf

- Berners-Lee, T. (2006) Linked Data. W3C. https://www.w3.org/DesignIssues/LinkedData.html

- Best Practices for Publishing Linked Data" (https://www.w3.org/TR/ld-bp/)

- Hooland, S. van, & Verborgh, R. (2014). Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. London: Facet Publishing.

- Hyland, B., Atemezing, G., & Villazón-Terrazas, B. (2014). Best Practices for Publishing Linked Data. Retrieved 15 May 2020, from https://www.w3.org/TR/ld-bp

- Knight, S. A., & Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. Informing Science, 8. http://inform.nu/Articles/Vol8/v8p159-172Knig.pdf

- Nousak, P., & Phelps, R. (2007). A Scorecard approach to improving Data Quality 2007. SUGI 27: Data Warehousing and Enterprise Solutions, PWC Consulting https://support.sas.com/resources/papers/proceedings/proceedings/sugi27/p158-27.pdf

- Platform Linked Data Netherlands. (2019). Publishing Linked Data in 9 steps. https://www.pldn.nl/wiki/BoekTNO/stappenplan

- Siebes, R., Coen, G., Gregory, K., & Scharnhorst, A. (2019). Top 10 FAIR Data & Software Things: Linked Open Data. Retrieved 15 May 2020, from https://librarycarpentry.org/Top-10-FAIR/2019/09/05/linked-open-data/

- Verborgh, R. & van Hooland, S. (2014). Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata. ALA Editions.

11. Appendix

CERL's Resources as Linked Open Data

The database and its contents. The Consortium of European Research Libraries (CERL) hosts a number of databases on Early Modern book history, including the Heritage of the Printed Book Database (HPB)[22], Material Evidence in Incunabula (MEI)[23], and a number of smaller, more specialized databases for research projects. As a central resource, both for use in CERL's own databases and beyond, it also maintains the CERL Thesaurus (CT)[24], a database that collects corporate and personal names, as well as associated names of printing places and printers, primarily from imprints of Early Modern books. The CT currently holds 1.383.482 records[25]. We are continuously updating and adding records, as well as having a small team of editors checking potential duplicates and merging them. Beyond our own interface, the CT offers its records in a variety of export formats (RDF/XML, RDF/Turtle, JSON-LD, JSON, YAML, UNIMARC). Internally, CT records are being stored as JSON objects in a CouchDB, with search realized through an ElasticSearch index. The interface which is developed and maintained in-house provides functionality for both end-users (search & display) and editors (record creation & editing).

Linked Open Data and the CT. Publishing the CT as Linked Open Data has never been a single, one-off project. Rather, we place emphasis on LOD as part of a continuous process of providing access to our data in accordance with current standards. The CT was established in 1999, with the technological limitations of its time, but has always been aligned with the ideals of openness and connectedness of LOD. Records are licenced under the Etalab Open Licence, which is roughly equivalent to ODC-BY or CC-BY 2.0[26]. The internal format of the CT underwent migration from a UNIMARC-derived format to JSON around 2018, but a first RDF representation of the data in RDF/XML serialization was already made available in 2012. Around the same time, we also began including links to other data providers in our records, starting with the German National Library's Gemeinsame Normdatei (GND), and adding more as libraries and other institutions began making their data available as LOD. Our most recent effort has been an exchange of data with Wikidata, adding both Wikidata identifiers to our records and providing assistance with adding CT identifiers to Wikidata. Many of these efforts occur in the course of our regular maintenance and development of the

[22] Online at <http://hpb.cerl.org>
[23] Online at <http://data.cerl.org/mei/>
[24] Online at <http://data.cerl.org/thesaurus/>
[25] These are divided into categories as follows: 26.623 corporate names, 35.816 place names, 89.994 printers, and 1.231.044 personal names.
[26] The licence is available under: <http://ddata.over-blog.com/xxxyyy/4/37/99/26/licence/Licence-Ouverte-Open-Licence-ENG.pdf>

CT, e.g. when ingesting data from other authority files, rather than in separate LOD-oriented projects.

How our data is being used. Beyond the day-to-day use of our databases through their own interfaces by the bibliographic research community, CERL's datasets are already being used in a variety of Digital Humanities projects. Most prominent is the CERL Thesaurus, which is used by multiple projects to normalize, e.g., printers' names[27].

Linking and contributing one's own records to the Thesaurus is also seen as a way of increasing their visibility and moving towards a Linked Open Data environment[28], while other researchers point out the ways in which the Thesaurus is already linked to their own database through the use of shared resources like the Gemeinsame Normdatei[29]. Beyond the Thesaurus, it is especially the Heritage of the Printed Book database that has become a focus of attention for bibliographic data science and metadata quality analysis[30]. This also motivates the focus of our efforts on these two projects which have already proven their uptake by the community.

Future directions. One major milestone for many LOD projects that we have not yet tackled is the provision of an endpoint for SPARQL queries over the published graph. While no such feature currently exists in our technology stack, we have recently started exploring the potential of adding an Apache Fuseki triple store, in order to enhance search capabilities and adding visualization options based on SPARQL queries[31]. We are planning to integrate this technology, and the lessons learned from building a prototype for an RDF-based interface to the HPB, with the CT in the context of an externally funded project over the next two years.

Another topic is the accessibility of LOD resources for different audiences. CERL has a large audience of domain experts who do not necessarily have the technological capabilities for working with tools like SPARQL in their home institutions but are nevertheless interested in the possibilities these technologies open up for their research. For this reason, we are exploring options for collecting, creating, and sharing scripts and other resources that

[27] Tuppen, S., Rose, S. & L. Drosopoulou (2016), Library catalogue records as a research resource: Introducing 'A Big Data History of Music', In: Fontes Artis Musicae, Vol. 63(2), 67-88; Montoya, A.C. (2018), The MEDIATE Project, In: Jaarboek voor Nederlandse Boekgeschiedenis 25.2018, 229-32; Tuominen, J. et al. (2018), Reassembling the Republic of Letters – A Linked Data Approach, Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, CEUR Workshop Proceedings; Dittmar, J. & S. Seabold (2019), New media and competition: printing and Europe's transformation after Gutenberg. CEP Discussion Papers (1600), Centre for Economic Performance, LSE, London.
[28] Van Egmond, M. (2016), Mapping early Utrecht printers and publishers: experiences with building a geographical interface, In: e-Perimetron, Vol. 11(4), 170-182; Kräutli, F. & M. Valleriani (2018), CorpusTracer: A CIDOC database for tracing knowledge networks, In: Digital Scholarship in the Humanities, Vol. 33(2), 336-46; Verger, N. (2018), The Printers' Devices database of the University of Barcelona. A resource for the study of printers' devices. Talk given at: Typography, illustration and ornamentation in the Early Modern Iberian Book World, 1450-1800, May 24-25, Marsh's Library, Dublin.
[29] Reinert, M. & D. Scholz (2017), Digital Humanities and the "Deutsche Biographie" as historical biographical information system. DH. Opportunities and Risks. Connecting Libraries and Research, Aug 2017, Berlin.
[30] Lahti et al. (2019), Bibliographic Data Science and the History of the Book (c. 1500-1800), In: Cataloging & Classification Quarterly, Vol. 57(1), 5-23; Király, P. (2019), Measuring Metadata Quality, PhD Thesis, University of Göttingen.
[31] See Walker, A. 2019. Improving access to bibliographic data. Representing CERL's Heritage of the Printed Book database as Linked Open Data. MA Thesis, Humboldt University. <http://wwwuser.gwdg.de/~walker5/docs/201905_thesis.pdf>

work with our data, e.g. in the form of Jupyter notebooks or modifiable SPARQL queries.

Challenges and recommendations.

Finding an ontology. Finding an appropriate vocabulary to express our data in was one of the major issues in implementing the first RDF export. However, this is not a one-off choice that never needs to be revisited: which standards survive and thrive is only obvious in hindsight, and new developments may lead to standards not available at the time. Our RDF export, mostly based on the RDA vocabulary, will probably need to be revisited in the near future, as ontologies – like technologies – age and develop. For the HPB, we are planning to adopt the emerging BIBFRAME.

Continuity. We see making our data accessible through current standards and technologies as part of our daily work of database enrichment, maintenance and migration – there is no clear separation between these and the publication as Linked Open Data. RDF export was added when it became a feasible and interesting option, and now that we are seeing the technology develop and mature, we continually assess our researchers' needs and iterate based on those and the options available to us. We are very invested in the long-term sustainability of CERL's resources and aim to ensure that our technology stack is stable and can be supported by our team in the long term.

Accessibility to non-technical audiences. We aim to make graph-based representations an additional access point to our data, but also retain an interface that is easy to work with and integrates well with the workflows of our users, who are largely domain experts without IT training.

LIBER

www.libereurope.eu