Linked Data at the Swiss Federal Archives

Status Report*

Dr. Cochard, Jean-Luc

[1]Swiss Federal Archives, Archivstrasse 24, 3003 Bern, Switzerland

Abstract. Linked Data is attracting increasing interest from the Swiss public administration. The Swiss Federal Archives are playing a leading role in this respect by investing significantly in the deployment of an infrastructure for publishing data in LD. This approach has enabled the institution to acquire in-depth knowledge on the subject and to consider integrating LD into its core applications and into the services it offers to the public.

Keywords: Linked Data, RDF, triplestore, Archival Information System, Database

1 Historical background

The Swiss Federal Archives (SFA) have been interested in Linked Data (LD) technology for almost 10 years now. Initially, a few studies were commissioned from academic institutions in order to allow the leadership of the SFA to get an overall idea of this technology, which was already of great interest in the field of libraries [1].

The interest in LD was further increased when the SFA took over a project for the publication of open government data (OGD) in the Swiss federal administration in 2013. In Sir Tim Berners Lee's 5-star model [2] LD is considered to be the optimal format for publishing data.

In 2014, a pilot infrastructure for hosting Linked Data was set up within the federal administration. The solution called LINDAS for Linked Data Service, has enabled various administrations to experiment with both data conversion and access to LD from web applications.

LINDAS also favoured the setting up in 2014 of a collaboration within Swiss archival institutions with the name aLOD (archival Linked Open Data) [3]. Its ambition was to concretely experiment with the conversion into LD of descriptive metadata managed by the AIS[1] of several Swiss institutions in order to gain experience in this field.

Between 2017 and 2020, the focus was on improving LINDAS in order to transform the prototype solution into a productive, reliable infrastructure capable of hosting large volumes of data. In parallel, additional studies were conducted to determine whether LD and the triplestore databases were able to meet the technical requirements of an AIS.

[1] Archival Information System

2 aLOD related activities

The archives participating in aLOD activities have set themselves the following goals:

1. To examine the opportunities for LD to achieve the mission of archival institutions.
2. Based on real descriptive metadata sets from their respective AIS, to transform and unify these metadata into LD datasets.
3. In doing so, to formulate "best practices" for transforming existing inventories (metadata) into LD.
4. To communicate and disseminate the achievements of the project within the archival community and the internal users, but also beyond, for example within the community of researchers in digital humanities, and also to exchange with the actors who contribute to the implementation of LD technologies (GLAM and beyond).
5. Demonstrate the potential for third party reuse of descriptive metadata in LD when made freely available (OGD), as for example in the context of hackathons on cultural data.

The different archives exported data from their AIS in the form of csv files that were then converted into LD using an ad hoc data model, since the RiC-O data model [4] was not yet available when this work was undertaken. Several particularities had to be taken into account in order to achieve a commonality of these data:

- The contents of the inventories had very different levels of detail from one institution to another. The data model therefore had to be enriched as new datasets were integrated.
- The language of these contents was different: French or German in this case. Fortunately, this aspect is well handled with language tags in RDF[2].
- The dates had variable numerical formats or even were in textual form. This is one of the aspects that took the longest time to be dealt with, without resulting in a clean and reusable solution.
- For each institution, a data export procedure had to be put in place. Even for institutions that used the same AIS, it was not possible to have a generic solution, as the content structures were quite different.

The data conversion produces triples like those associated with the AFS record with the signature "B0#1000/1483#3792*" (see Fig. 1). Thanks to this uniform representation of the data from the different archives and to the fact that these Linked Data are directly accessible on the web via LINDAS and its SPARQL interface, it has been possible to have an experimental prototype for the visualisation of all these data (see Fig. **2**). This representation includes a histogram of the number of records per date, which is unusual in archival web portals but could be useful to identify the density of information over time on a specific subject.

[2] Resource Description Framework: a formal model to define graph structures.

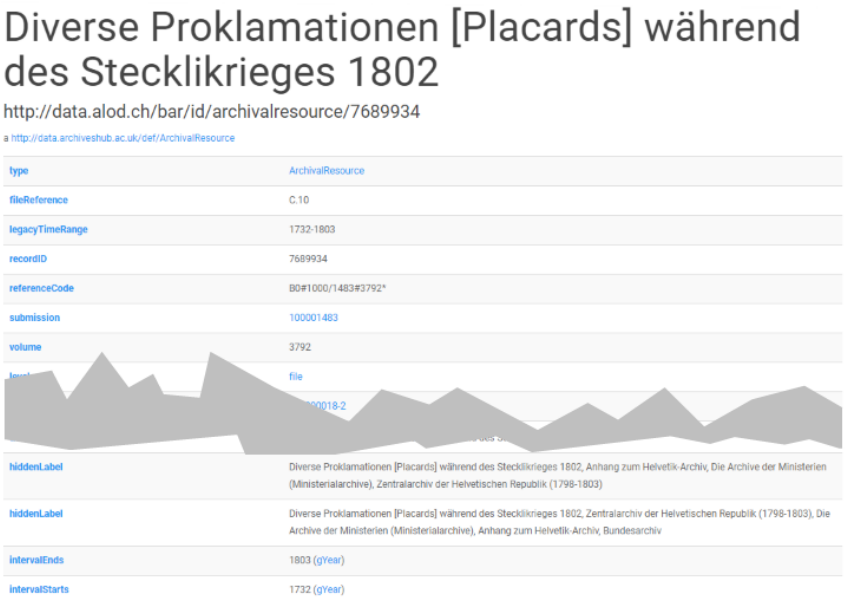

Fig. 1. Example of an entity from the AIS of the AFS, of type "File", converted to LD with the ad hoc data model used in 2015.

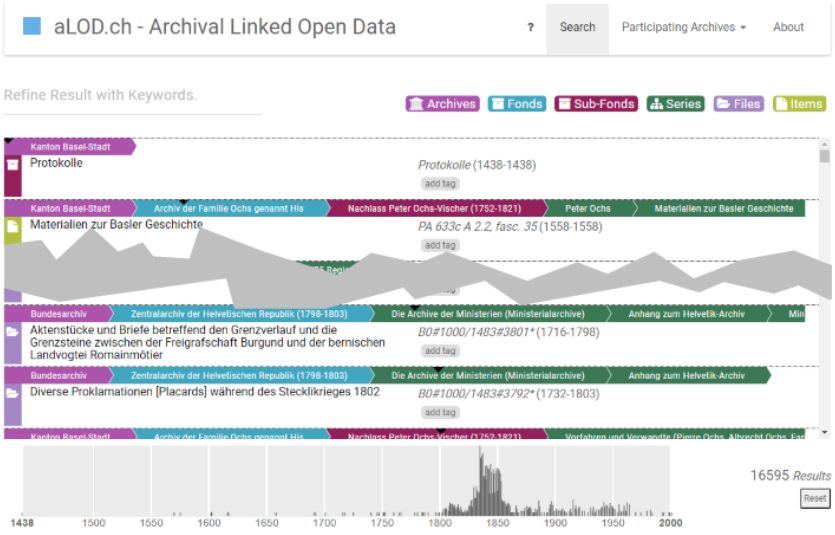

Fig. 2. Screenshot of the experimental application representing descriptive metadata of institutions participating in aLOD.

3 LINDAS

LINDAS as a linked data hosting infrastructure has been enhanced since its first release to become a productive infrastructure. This enhancement was carried out between 2017 and 2020. Its general structure is described schematically below (see Fig. 3). In the centre, there are several triplestores to allow testing, integration and finally production of new datasets. Data conversion can be a recurring or a one-off process. In any case, an ETL pipeline is implemented, the execution of which can be scheduled according to the updating of the source data.

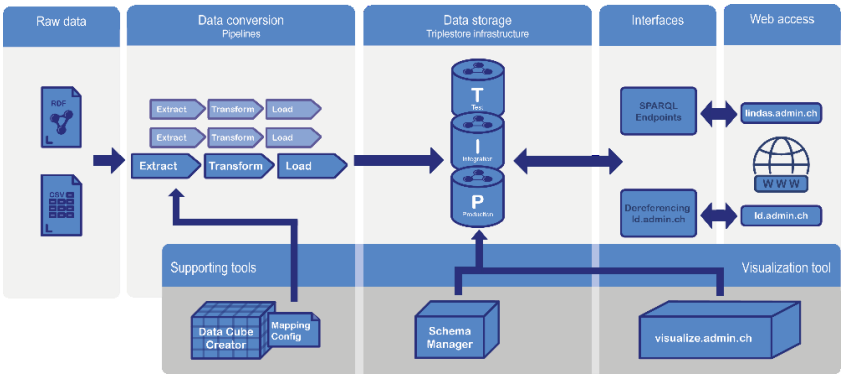

Fig. 3. Structure of LINDAS with its satellite solutions.

To ease the definition of these conversion processes, the Data Cube Creator tool, specialised in the conversion of OLAP cubes [5], has been implemented. This tool allows configuring the conversion of this type of data without having in-depth knowledge of the W3C cube model [6], used for this purpose. This auxiliary solution allows many administrations to publish data in LD as LOGD. In addition, to enrich data documentation, the Schema Manager tool allows the modelling and publication of schemas and ontologies [7]. This is central to the long-term archiving of LD, as the description of the modelling schemas is as important as the data itself to define the semantics of a dataset.

To complete the infrastructure, the graphical visualisation of data hosted in LINDAS can be parameterised using the Visualize tool [8]. This solution works as an accelerator for the adoption of LD as it facilitates the production of interactive graphical representation in web pages or digital reports if the data is first converted to LD and published in LINDAS.

4 Linked Data as core technology of an AIS

We believe that LD is the optimal solution for publishing data and making it accessible on the web. The question we asked ourselves in relation to our core activities as an archive is whether LD and more specifically the RDF model could be used as the central

database of an AIS. In 2018 and 2019, two studies[3] were conducted by research institutes to verify certain aspects of this technology in relation to our own issues.

It appeared from this work that there are suppliers of triplestores able to deliver solutions that are perfectly suited to our needs. Thus, Stardog version 5.2 allowed us to build a graph of 10B triples by reading files of 100M triples with an average and stable execution time of 20 min. per file. This amount of data is much more than what we estimate we will eventually have to manage in our AIS: 100-500M triples.

Updates are crucial operations that are implemented by Delete and Insert functions. In our test, 1M updates were performed in 12 sec. on average. And finally SPARQL queries of different complexity combined with Insert, all at different frequencies, have had sub-second response times.

Therefore, we are confident that, if the triplestore is installed on suitable servers, this technology will be performing well as the core database of an AIS.

Another issue that has been studied is whether the RDF model is as powerful as Property Graphs (e.g. Neo4j [9]). Fortunately the evolution of RDF to RDF-star [10] and its pendant SPARQL to SPARQL-star considerably reduces the expressive advantage of Property Graphs while maintaining the advantage of RDF which is a W3C open standard. As such, the RiC-O standard is written in RDF but is designed to evolve quickly into RDF-star when this new standard is approved.

In our opinion, there is no reason why an AIS should not be developed with a triplestore at its core as a central database.

5 Future developments

If LD can be implemented at the core of an AIS, it can also have other roles. Here are two areas we are considering working on in the coming years.

5.1 Publication of database content

By publishing datasets in LD, public administrations take a first step towards publishing entire databases for public reuse. However, archives also hold databases in their archive holdings, ideally in SIARD format [11]. Unfortunately, this format is not designed for web publication of data and their structure. A conversion from SIARD to LD seems to be a promising and feasible way to fill this gap.

5.2 Testing RiC-O

RiC-O in its current version 0.2 is a very promising proposal that still needs to be tested in the very different contexts of Swiss archives. To this end, LINDAS and its data conversion environment will allow us to test the conversion of the descriptive metadata of

[3] These reports have not been published but the author can provide you with a copy if desired.

our inventories according to the RiC-O standard. Only then will it be possible to identify possible gaps in the model and to establish best practices in the way of proceeding with this conversion task.

References

1. Godby, Carol Jean.: The Relationship between BIBFRAME and OCLC's Linked-Data Model of Bibliographic Description: A Working Paper. Dublin, Ohio: OCLC Research (2013), https://www.oclc.org/content/dam/research/publications/library/2013/2013-05.pdf, last accessed 2021/07/02.
2. 5-Star Open Data, https://5stardata.info/en, last accessed 2021/07/02
3. aLOD homepage, http://www.alod.ch, last accessed 2021/07/02
4. RiC-O Version 0.2 homepage, https://www.ica.org/standards/RiC/RiC-O_v0-2.html, last accessed 2021/07/02
5. OLAP Cube homepage, https://en.wikipedia.org/wiki/OLAP_cube, last accessed 2021/07/04
6. The RDF Data Cube Vocabulary, https://www.w3.org/TR/vocab-data-cube/, last accessed 2021/07/04
7. Zazuko Ontology Manager, https://zazuko.com/products/ontology-manager/, last accessed 2021/07/04
8. Visualize homepage, https://www.visualize.admin.ch/en, last accessed 2021/07/04
9. Neo4j homepage, https://neo4j.com/, last accessed 2021/07/04
10. RDF-star and SPARQL-star Community Group Report, https://w3c.github.io/rdf-star/cg-spec/editors_draft.html, last accessed 2021/07/04
11. SIARD Suite homepage, https://github.com/sfa-siard, last accessed 2021/07/04