Linked Data on historical persons: publishability, interconnectivity and sustainability

Leon van Wissen, University of Amsterdam (UvA)
Richard Zijdeman, International Institute of Social History (IISG)
Rick Mourits, Radboud University (RU)
Ivo Zandhuis, International Institute of Social History (IISG)
Lodewijk Petram, Huygens ING
Laura Hollink, CWI Amsterdam

Type
Panel

Keywords
Linked Open Data, Vocabularies, Infrastructure, Data Reuse, RDF, Interdisciplinary Datasets

Introduction

We have come a long way since the floppy-drive based datasets researchers would refuse to share. Actually, quite a lot of data is coming out as Linked (Open) Data, but this presents us with new challenges. Specifically for the case of **publishing, connecting and sustaining Linked Data on persons** we pose five challenges that we would like to discuss with the community, especially now that the problems we face is becoming more widespread, as more and more cultural heritage institutions open up their collections, and digitization initiatives of archives take flight.[1]

Challenges

Published Linked Data

The first two issues concern *publishing* Linked Data ourselves.

1. **How can we model our datasets in such a way that they integrate well with existing Linked Data? Which are the (preferred) vocabularies or modelling strategies to choose from? And where can these be found and/or published once created?**
   This is very much centered on using RDF as a means to achieve interoperability, but what if vocabularies are not yet published as RDF? Too often, a project creates its own ontology or model, without providing proper documentation. Having a hub, where one can see commonly used schemas or recommended modelling strategies in Benelux projects would

[1] E.g. through the READ project (https://eadh.org/projects/read) that opens up handwritten documents.

be helpful, especially if they are supplemented with real data examples.[2] Second, having a common hub for vocabulary mappings would also increase dataset interoperability.[3]

2. **Where do we make our datasets accessible? Where can we host our Linked Data so that it can be accessed by others interactively?**
   There is an increasing trend in opening up project data from the project's start. But, it is hard to publish your data as Linked Open Data that is both findable and reusable with the help of stable permalinks (i.e. URIs). Reasons for this are a lack of proper infrastructure that allows for curation as well as publication, and a lack of triplestores where data can be hosted.

   Opening up your data at an early stage possibly prevents other projects from doing duplicate work. Moreover, it enables others to provide feedback on the project, when it is still shapeable. It is thereby essential to have community serving and stable URIs others can rely on. Unfortunately, it is not a given that URIs stay permanent, are dereferenceable, and are meaningfully described. The adoption of a shared permalink strategy and/or service that is accessible for all research projects may be able to assist in this matter.[4]

Interactive Linked Data

Once a dataset is published, we would like to advocate its existence in order for others to benefit from it and overcome the challenge of *lack of interactivity* of Linked Data.

3. **How can we make others interact and reuse our data on the level of vocabulary, thesauri, or resources?**
   Data analysed, enriched, and (re)modelled in a research project is often focussed at a particular time period, has georeferencial restrictions, or is aimed at a specific part of the population. As such, it is kept in one data silo, ideally, where possible, linked to common thesauri such as ULAN or VIAF (or Dutch equivalents RKDArtists and the Dutch Thesaurus of Authors (NTA)). Logically, such outgoing links can only be made to authoritative records for the entities that are already known somewhere, and that are already well-documented. This implies even beforehand that the entities that do not conform to such criteria are unlinkable, or are hindered in being spotted by similar and/or follow-up projects. The question is then how to disclose an entity when an authoritative record is not available?

   We have to face the fact that we cannot construct authoritative databases of the same level and quality as the mentioned (curated) person thesauri. However, it is our responsibility to

---

[2] This can for instance be the recommendation to model person names in the Person Name Vocabulary (https://w3id.org/pnv), and/or to choose an event-based modelling strategy. A list with more commonly used vocabularies is kept in the 'Awesome Ontologies for Digital Humanities' list on GitHub (https://github.com/CLARIAH/awesome-humanities-ontologies).

[3] There are plans to develop a 'CLARIAH Vocabulary Registry' by the CLARIAH IG-Vocabularies (https://github.com/CLARIAH/IG-Vocabularies).

[4] Services such Linked Open Vocabularies (https://lov.linkeddata.es/dataset/lov/) and w3id.org (https://w3id.org) come a long way, but the diversity of vocabularies they offer hamper homogeneous use of URIs and vocabularies.

also keep an eye for the under-represented and at least not go along in any existent bias that is present in biographical databases.[5] Ideally, you need qualitative research to come to an authority record. But what if this is done through automatic linkages, clustering techniques and/or other automatic methods used in a research project? How can this data be used next to authoritative records?

4. **How can we construct links between datasets generated in research projects, on the level of individual resources?**
   A first step in connecting two datasets is to link on the level of the constructed resources, for instance stating that one person is the same as another. This link, which is likely part of a specific linkset, has to be stored somewhere with sufficient provenance information.[6] Being able to make and publish such links bridges the information in two datasets that for instance vary in scope, and increases the informational value of both datasets.

   At this moment we feel that there is no infrastructure to store and curate such links, independent of the datasets they apply to. Wikidata comes close, as a data hub for outgoing identity links, but is not suitable for usage with primary sources, nor is it viable to store all 'non-notable' entities in Wikidata. We need a place to store these links, including the evidence on which they are based. In that way, one is not limited to the level of the constructed entities alone, but one can also make use of the dataset's internal logic and original evidence (i.e. primary sources). Again, having information on the provenance and uncertainties surrounding entities is crucial to make this work.

Sustainable Linked Data

We finally address a challenge regarding the *sustainability* of Linked Data, which is especially related to the fact that Linked Data often comes forth from research projects.

5. **Where are datasets that are created in the time span of a research project hosted, curated, and kept alive/accessible *after* the project finishes?**
   What happens with the data when its publishing authority is no longer available or does not have the funds anymore to work on the data. The fact is that by creating a URI, you automatically attribute an authority (cf. namespace) to your entities. To facilitate follow-up projects and the semantic web in general, these datasets should be kept alive somewhere, by at least making them browsable and queryable (e.g. in a triplestore), and thereby interactive. Depositing your data into cold storage at an archiving repository does not suffice, as the individual resources inside the data will not be accessible in the LOD-cloud anymore, hence the benefit of Linked Data as RDF is annulled.

---

[5] Think of a gender and ethnic bias.

[6] That is, in the same way individual datasets are, or being part of a dataset. The requirement of knowing the authority of this data (i.e. who made it and with what intention/plan) applies equally to both linksets and datasets.