Improving interoperability after IDS: Connecting civil registry data with open vocabularies

1. Introduction

Over the past decades, a wide range of historical demographic datasets have become available. Scholars from Asia, Canada, Europe, South Africa, and the US are well aware of each other's efforts, as the SSHA and other conferences have been excellent platforms to share good practices and build towards better datasets. As a result, researchers from a wide array of countries now have access to high-quality datasets on individual life courses and family connections.

The wider availability of international databases has been a huge gain and allows for exciting comparisons between contexts. Yet, comparative research has been limited by the fractured nature in which the datasets have been built. Datasets were built at different times, on different sources, by different people who spoke a variety of languages, stored at different locations, and are offered via different methods of retrieval. Thus, the structure in which data is provided has been very different: file formats differ, variable names are different, and categories are unstructured. As a result, research studies often focus on one particular database, hampering comparative research.

To optimize our infrastructure for comparative research, our databases need to become Interoperable. Therefore, the IDS has been developed, so that information from each different database can be exchanged in a similar data format. This made it possible to compare databases on a larger scale (see e.g. Quaranta & Sommerseth, 2018). However, to fully solve issues with Interoperability and Reusability, information from the variables themselves needs to be standardized and made available in an environment that is as FAIR as possible.

2. How to learn from each other's insights?

One of the biggest challenges in the transition to *open science* is making data interoperable. Without coordination, database managers tend to come up with different descriptions for the same information. To tackle this problem, vocabularies and ontologies have been designed to standardize how data in datasets is being described. Sometimes these standardization efforts are very straightforward and apply to very broad contexts, whereas others are of general use to specific communities. For historical data, however, most of these standardization efforts are problematic as they were made to describe contemporary data and underappreciate how information and meaning can change over time. For example, places and their names change over time, occupations and social standing shift, and causes of death have different meanings between contexts. Existing vocabularies standardize these historical data at the cost of losing or misinterpreting information, which is why multiple historical demographers developed their own ontologies.

Historical demographers from a wide array of countries have built databases to reconstruct the lives of people in Europe, North America, and East Asia. The ontologies of these databases were designed to "stay true to the source", so that datasets have sophisticated designs to model local peculiarities and changes in meaning over time. Each of these local efforts has made it possible to standardize defunct phenomena, historical distinctions, and general changes over time – though only within the geographic scope of their projects. Each of these standardization schemes is worth its weight in gold, as it unlocks a wealth of historical data and contains years of insight in the historical sources and context. However, there is no clear overview of the ontologies and vocabularies in historical demography.

The coming year, we would like to take the first steps towards a common language by gathering and sharing the different vocabularies and ontologies in the field. Collecting this information requires a small team that knows the field well, has expertise in presenting data, and has time to invest in ontology design. We will gather the vocabularies that historical demographers currently use to standardize their data, map the relationships between them, and publish the results on a webpage at the IISG data hub, so that everyone in the field can easily find and access the existing ontologies/vocabularies and see how they relate. By gathering and sharing the ontologies, historical demographers can learn from each other's insights, prevent the re-invention of vocabularies, and ensure that data is interoperable. But most importantly, it lays the groundwork for a move towards open data in historical demography, as common ontologies allow for general-purpose software, make replication studies easier, and are the steppingstone to Linked Open Data.

3. Linked Data: Making data even FAIRer

More than a decade ago, Berners-Lee presented his view on the next web: the web of data. A Hypertext Transfer Protocal (HTTP) was made to communicate text and exchange documents smoothly between different computers. This has resulted in the internet, which made information much easier to share. However, there is still "a lot of untapped potential" in the internet, as data within documents is still not easily findable, accessible, interoperable, or reusable. Therefore, Berners-Lee opted to use the principles behind HTTP on data, which he based on three principles:

1. Each data point/observation gets its own address/Unique Resource Identifier (URI), so that it is retrievable online.
2. Data returns in a standard format, so that information is interoperable.
3. Relations between data points/observations are described with URIs, so that we know what the relation between different data points is. These relations also get their own URI.

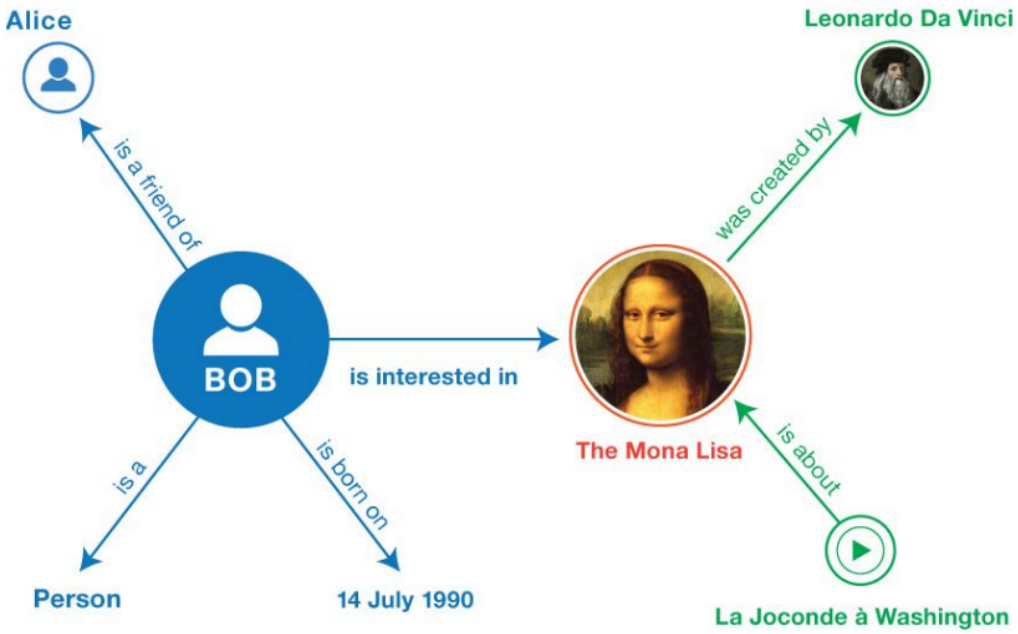

Figure 1: An example of linked data. Data points/observations are connected using a description. In databases these descriptions are put online as a unique HTTP-address.

Today, Berners-Lee's call for Linked Data has resulted in numerous 'end points' that allow for live querying of data of any kind. In his vision, Berners-Lee calls for 5 star linked data, which is made available: 1) online, 2) as structured data, 3) in an open format, 4) with standardised URIs, 5) and connected to other linked data. Most large databases already provide their data with three stars: on the web, machine readable and in a non-proprietary format. However, the key point of Linked Data is to use and reuse common vocabularies to describe data, adding the fourth star. The fifth star is earned by providing the data in the Resource Description Framework (RDF), a W3C standard for data exchange. The latter also allows for data to be presented in a graph-like structure, rather than a tabular structure. This is important, as graphs add descriptions between variables. With data in RDF format, different coding schemes can therefore be applied at the same time, so that researchers can relatively easily switch between different coding schemes.

4. So where to go from here?

To start cooperating, we first need to get an overview of how the existing historical databases encode their data. We recently received an NWO Open Science Fund to gather the different ontologies and vocabularies that have been designed by these institutions. To get a broad outline of the existing vocabularies we would like to first collect the coding schemes used by the 8 bigger data centers[1] and identify the connections between them. These 8 data centers are a logical place to start, as they have the most developed infrastructure and are important regional hubs in historical demography. This will give us a feeling for how the ontologies in the field were designed and how much they differ from one another. Moreover, it gives us the opportunity to map the interoperability between the different coding schemes and indicate where ontologies and vocabularies overlap or are complementary to one another.

Based on the input from the audience and a second round of input by email, we supplement our overview of ontologies and vocabularies. For each of these datasets, we will map the compatibility with the ontologies and vocabularies of the 8 bigger institutions. In March 2022, we will present the updated list to the field at the European Society for Historical Demography conference (ESHD), so that we can receive an extra round of feedback and finalize our overview of available ontologies and vocabularies in the field. With this input we will design a website that is easily perusable for fellow historical demographers. This list of ontologies and vocabularies will be a vital steppingstone for general-purpose software, replication studies, and Linked Open Data.

Ideally, our insights can then be queried with Linked Open Data (LOD). LOD itself ticks a lot of the FAIR boxes, but fails when it comes to its programming language SPARQL, a very popular method to request and reuse data from the LOD cloud. In contrast to Linked Data, these queries are often not stored and shared with other users, as there is no proper way of doing so. As a result, queries are 'hidden' in application code or in files on hard drives. Even when shared online, for example on platforms like GitHub, the lack of metadata standards for queries make them hard to find and reuse, hampering the reproducibility of scientific results based on LOD. To overcome this obstacle, we will create a connection between two proven open source solutions, Yasgui and grlc (*'garlic'*), to enhance the Findability and Reusability of SPARQL queries and make Linked Data more user-friendly. Moreover, we will build a collaborative query editing feature in the SPAQRL editor Yasgui, allowing for more and less experienced users of linked Data to collaborate.

[1] Barcelona, Hong Kong, Minnesota, Tromsö, Budapest/Graz, Montreal, Salt Lake City, Lund, and ourselves.

Chronicling our coding schemes and offering a platform where we can share our code will be a huge boost to further cooperation within the field of historical demography. But, perhaps more importantly, we believe the proposed ideas will have a large impact on the Open Science practice of the LOD community. The Yasgui service is widely used across scientific domains and beyond, e.g. serving as frontend to the European Data Portal. grlc has 4,948 unique users in its public instance on Github since July 2016, amongst which various large companies (Elsevier, TNO, NewGen Chennai). We therefore expect that the feature to store queries via Yasgui in grlc format will have impact across all academic disciplines using LOD, as well as throughout various industries. It also allows communities around LOD clouds, such as Wikidata, to offer their SPARQL queries FAIR-ly. Finally, based on our teaching experience in various disciplines, we believe this will aid students in mastering SPARQL as they learn efficiently via example queries. The enhanced findability of queries and collaborative query editing will provide them more examples to learn from.

5. Conclusion: We have an ambitious year ahead of us

Rick

- Sent out surveys to the bigger data centers
- Have an informal online meeting on research progress
- Go for a second round in March
- Present a website in August/September 2022

Richard

- Provide tools for efficient data querying
- Present a demo at the next SSHA