Common Language for Accessibility, Interoperability, and Reusability in Historical Demography

Rick J. Mourits[1 [0000−0002−2267−1679]], Tim Riswick[2 [0000−0003−1401−6284]], and Rombert J. Stapel[3 [0000−0001−6394−260X]]

[1] International Institute of Social History, Amsterdam, the Netherlands
rick.mourits@iisg.nl
[2] Radboud University, Nijmegen, the Netherlands
tim.riswick@ru.nl
[3] International Institute of Social History, Amsterdam, the Netherlands
rombert.stapel@ru.nl

Abstract. One of the biggest challenges in the transition to open science is making data interoperable. Normally, ontologies and vocabularies are used to describe data, but these are generally problematic for historians as existing ontologies and vocabularies are insensitive to temporal variations. Within history, the subdiscipline of historical demography is a forerunner in dealing with this problem, as it studies large-scale reconstructions of populations and life courses. Historical demographers have designed their own ontologies and vocabularies to standardize historical data. We gathered these schemes to create an overview, so that we can standardize existing insights into a common language for historical (demographic) data.

Keywords: Interoperability · Vocabularies · Historical Demography

1 CLAIR-HD

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 field agrees that a common language is necessary to make historical demographic databases FAIR. 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 gathered the vocabularies that historical demographers currently use to standardize their data, mapped the relationships between them, and will publish the results on a webpage, 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 groundworks 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.

2 Methods

Information on the existing vocabularies and ontologies was gathered in multiple rounds. Our initial goal was to get a broad outline of the existing vocabularies. Therefore, we have contacted the bigger data centers in Asia, Europe, and North America. These data centers were a logical place to start, as they have the most developed infrastructure and are important regional hubs in historical demography. We gathered information on the different ontologies and vocabularies that have been designed by these institutions. This gave us a feeling for how the ontologies in the field were designed and how much they differ from one another. Moreover, it gave us the opportunity to map where ontologies and vocabularies overlap or are complementary to one another.

3 Results

Our initial goal was to get a broad outline of the existing vocabularies and show the overlap between them. Once the data came in, it became clear that information in the field was less standardized than we expected based on preliminary enquiries before the project started.

Generally, occupational data was the most standardized as occupational status schemes have been in place for several decennia. When analyzing or coding occupational information, database managers and researchers can choose from a list of fully standardized occupational titles, groups, class, or status measurements. Variables such as religion and relations were much less standardized. In these cases, a standard vocabulary is either not available or accepted within the field, so that database managers and researchers often use their own categories. Causes of death provided an interesting case as a large group of international scholars is in the process of creating an approach to standardize and code them. To help the field in adopting open science practices, we decided to list per variable whether accepted vocabularies exist, how much support they have, and whether conversion tables exist to translate between rival encodings. Furthermore, we list whether teams in the field are working on new vocabularies and how these researchers can be contacted.

4 AISoLA

At the conference, we show a detailed overview of the different vocabularies and codings, such as:

- National collections of occupations titles, such as: [5], [7]
- HISCO [9], HISCLASS [10], HISCAM [4], SOCPO [11], OCC1950 [6], and other social status measurements
- ICD10h [3] to classify historical causes of death
- The Linked International Classification for Religions [2]

We also pay special attention to the schemas designed by A2A-LD, the Intermediate Data Structure [1], and MOSAIC [8] that allow users to model databases from different sources and countries with one model.

Acknowledgements This publication is part of the project "Common Language for Accessibility, Interoperability, and Reusability in Historical Demography (CLAIR-HD)" with project number 203.001.166 of the research programme Open Science Fund which is (partly) financed by the Dutch Research Council (NWO).

References

1. Alter, G., Mandemakers, K.: The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. Historical Life Course Studies **1**(1), 1-26 (2014).
2. Askhpour, A.: LICR. IISH Data Collection (2017). https://hdl.handle.net/10622/MHJWRZ
3. Janssens, A.: Constructing SHiP and an international historical coding system for causes of death. Historical Life Course Studies **10**, 64–70 (2021).

4. Lambert, P.S., Zijdeman, R.L., Van Leeuwen, M.H.D., Maas, I., Prandy, K.: The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. Historical Methods: A Journal of Quantitative and Interdisciplinary History, **46**2, 77–89 (2013).
5. Mandemakers, K. et al.: HSNDB Occupations. IISH Data Collection (2020). https://hdl.handle.net/10622/88ZXD8
6. Mourits, R.J.: HISCO-OCC1950 crosswalk. Dans Easy (2017). https://doi.org/10.17026/dans-zap-qxmc
7. Pedersen, B., Holsbø, E., Andersen, T., Shvetsov, N., Ravn, J., Sommerseth, H.L., Bongo, L.A.: Lessons learned developing and using a machine learning model to automatically transcribe 2.3 million handwritten occupation codes. arXiv preprint arXiv:2106.03996 (2020)
8. Szołtysek, M., Gruber, S.: Mosaic: Recovering surviving census records and reconstructing the familial history of Europe. The History of the Family, **21**1, 38–60 (2016).
9. Van Leeuwen, M.H.D., Maas, I., Miles, A.: HISCO: Historical international standard classification of occupations. Leuven University Press, Leuven (2002).
10. Van Leeuwen, M.H.D., Maas, I.: HISCLASS: A historical international social class scheme. Leuven University Press, Leuven (2011).
11. Van de Putte, B., Svensson, P.: Measuring social structure in a rural context Applying the SOCPO scheme to Scania, Sweden (17 (th)-20 (th) century). Belgisch Tijdschrift voor Nieuwste Geschiedenis-Revue Belge d'Histoire Contemporain **40**1-2, 249–293 (2010).