From Legacy Data to Intertwined Historical Census and Survey Data

A Case Study in Linked Research Data Modeling between Knowledge Organization and Knowledge Representation

Ingo FRANK

Library and Electronic Research Infrastructure Division, Leibniz Institute for East and Southeast European Studies, Regensburg, Germany

Context Although conceived early in the 1960s by Nelson [22] in his vision of a digital research environment for a historian, such hypertext-based research tools [e. g. 8] have not prevailed in the daily work of historians due to the problem of incomplete or inaccurately prepared sources [cf. 30]—i.e. the lack of adequately modeled research data. Nowadays, the linked data paradigm provides a solution for source-oriented [cf. [14] p. 190 f.] modeling and integration of historical sources as research data.

Objectives The paper shows how linked data technology can be applied to make research data FAIR [32] for research in digital history. Following the linked data approach [4], I focus on the preparation of legacy historical census and survey data for interoperability and the integration and distribution of the FAIRified data with historical maps and additional relevant geospatial data (e. g. the GREG ethnicity dataset [29] as a digital version of the Soviet Atlas Narodov Mira from 1964) in an institutional research data repository [see also [31]].

Methodology We use RDF for metadata as well as for data. In a pragmatic way we apply DCAT (Data Catalog Vocabulary) [1] and Disco (DDI-RDF Discovery Vocabulary) [6] to make research data findable. Historical census data were created as linked data from legacy Excel tables by Meroño-Peñuela et al. [21] and from tables in TEI markup by Bayerl and Granitzer [2]. We follow these projects in using the RDF Data Cube Vocabulary (DQ) to model statistical data [18]. In addition to DQ for modeling observations, we use Disco for documentation of variables in survey data. This improves the task of data review and the retrieval of relevant microdata [27]. Zapilko and Mathiak [33] p. 116] propose an improved linked open data workflow for gathering, cleaning, and harmonizing statistical data from different sources. We extend this proposed procedure to a research
data management workflow as a kind of FAIRification process selected legacy research data is prepared, enhanced, and harmonized in a case study focusing on modeling temporal coverage of research data in metadata and the data itself (i.e. dimensions in statistical data), interlinking historical places by using gazetteers and geographical coverage (meta)data in GeoSPARQL and WKT [see 19] and identifying historical agents and their roles by linking to authority files and encyclopedic knowledge in DBpedia or Wikidata and classification systems (e.g. HISCO). We reuse project specific coding schemes to integrate the data and to create the metadata to describe the datasets. The coding schemes are enhanced and modeled in SKOS as generic data model for knowledge organization systems. This enables the alignment with authority files, gazetteers, and knowledge organization systems in order to identify agents, places, time periods, and topics across datasets.

Implications Implications for research in computational history are besides the benefits of a common data model for statistical analysis [33, 20] new possibilities for analysis of integrated survey data (e.g. structural topic modeling of open-ended survey answers [25]). Based on more granular source-oriented modeling, the logical structure of questionnaires could be further exploited for advanced analysis [12]. Last but not least, digital source criticism [e.g. 7] could be supported by accessing integrated source material prepared in a shared data model.

Conclusions We designed an application profile [15] mainly based on DCAT including elements from Disco to describe methodological details of research data. We use DQ to move from metadata to model the data itself, i.e. to provide data distributions not only as CSV files, SPSS or Stata files etc., but also to represent the data as linked data. From a FRBR perspective, looking back to Nelson’s early hypertextual vision of intertwined research data, the complex interrelations between digitized source material, texts, research data, and knowledge organization systems can be precisely conceptualized in the conceptual framework of FRBR (see also Borgman [5] about the possible shortcomings of Nelson’s early view). By relating datasets and data distributions to FRBR’s so-called WEMI entities (Work, Expression, Manifestation, Item) via mapping to corresponding classes in FaBiO (FRBR-aligned Bibliographic Ontology) [24], we are able to improve versioning of research data and to enable more detailed citation of research data based on enhanced, i.e. more granular provenance information.

Outlook The thought experiment of the “Ideal Chronicle” [9] demonstrates the difficulties to get from data modeling to reality representation—i.e. modeling historical reality as conceptualized and represented in different interpretations of historical sources [see 13]. Therefore, we have to consider ontologies beyond mere data modeling and technical considerations [see also 17] which could be applied in a digital working environment or “contextualising knowledge system” like ResearchSpace [23] in order to further improve the interoperability and reusability of research data: CRMsci to model observations, CRMgeo [16] to represent places, and the CRM extension MIDM [26] to evaluate knowledge representation of multiple perspectives on historical phenomena based

---

1See FAIRification process scheme on the GO FAIR website: https://www.go-fair.org/fair-principles/fairification-process/

2See the latest linked data representation of HISCO (Historical International Standard of Classification of Occupations): https://druid.datalegend.net/HistoryOfWork/HISCO-latest
on different historical sources and their interpretations. Additionally, perspectives could be further specified [see [3] p. 29f.] by considering HiCO (Historical Context Ontology) [10] to represent interpretation acts carried out by different historians based on the same sources but achieving controversial conclusions about historical circumstances.

References


