Data for History 2020
Modelling Time, Places, Agents

From Legacy Data to Intertwingled 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[1]: 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[2]). 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 [28] perspective, looking back to Nelson's early hypertextual vision of intertwingled 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 (*W*ork, *E*xpression, *M*anifestation, *I*tem) 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 [11] extension MIDM [26] to evaluate knowledge representation of multiple perspectives on historical phenomena based

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

[2] See 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. 29 f.] 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

[1] Riccardo Albertoni et al. *Data Catalog Vocabulary (DCAT)*. Version 2. W3C Recommendation. Feb. 2020. URL: https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/.

[2] Sebastian Bayerl and Michael Granitzer. "Data-Transformation on Historical Data Using the RDF Data Cube Vocabulary". In: *Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business*. i-KNOW '15. Association for Computing Machinery, 2015.

[3] Klaus Bergmann. *Multiperspektivität: Geschichte selber denken*. 3rd ed. Methoden historischen Lernens. Schwalbach/Ts.: Wochenschau Verlag, 2016.

[4] Victor de Boer, Albert Meroño-Peñuela, and Niels Ockeloen. "Linked Data for Digital History: Lessons Learned from Three Case Studies". In: *Historiografía digital. Proyectos para almacenar y construir la Historia*. Ed. by Mirella Romero Recio and Jesús Colmenero Ruiz. Anejos de la Revista de Historiografía. Madrid, 2016.

[5] Christine L. Borgman. "Data, Metadata, and Ted". In: *Intertwingled: The Work and Influence of Ted Nelson*. Ed. by Douglas R. Dechow and Daniele C. Struppa. Cham: Springer International Publishing, 2015, pp. 67–74. URL: https://doi.org/10.1007/978-3-319-16925-5_10.

[6] Thomas Bosch et al. "DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data". In: *Proceedings of the WWW2013 Workshop on Linked Data on the Web*. Vol. 996. CEUR Workshop Proceedings. 2013. URL: http://ceur-ws.org/Vol-996/papers/ldow2013-paper-12.pdf.

[7] Benjamin Bridgman. "What Does the Atlas Narodov Mira Measure?" In: *Economics Bulletin* 10.6 (2008), pp. 1–8. URL: https://ideas.repec.org/a/ebl/ecbull/eb-08j10005.html.

[8] Frank Colson. "Case Study C: A Hypermedia Database Management System: Microcosm as a Research Tool". In: *Databases in Historical Research: Theory, Methods and Applications*. Palgrave Macmillan, 1996, pp. 69–72.

[9] Arthur Danto. *Narration and Knowledge*. Columbia University Press, 1985.

[10] Marilena Daquino and Francesca Tomasi. "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects". In: *MTSR 2015: Metadata and Semantics Research*. Vol. 544. Communications in Computer and Information Science. 2015, pp. 424–436.

[11] Martin Doerr. "The CIDOC CRM: An Ontological Approach to Semantic Interoperability of Metadata". In: *AI Magazine* 24.3 (2003), pp. 75–92.

[12] Griffith Feeney and Samuel Feeney. "On the logical structure of census and survey questionnaires". In: *Genus* 75.1 (2019).

[13] Ingo Frank. "Multi-Perspectival Representation of Historical Reality: Ontology-Based Modeling of Non-Common Conceptualizations". In: *Proceedings of the Joint Ontology Workshops 2019*. Vol. 2518. CEUR Workshop Proceedings. First International Workshop on Ontologies for Digital Humanities and their Social Analysis. 2019. URL: http://ceur-ws.org/Vol-2518/paper-WODHSA3.pdf.

[14] Charles Harvey and Jon Press. “Source-Oriented Database Systems”. In: *Databases in Historical Research: Theory, Methods and Applications*. Palgrave Macmillan, 1996. Chap. 7, pp. 190–217.

[15] Rachel Heery and Manjula Patel. “Application Profiles: Mixing and Matching Metadata Schemas”. In: *Ariadne* 25 (2000). URL: http://www.ariadne.ac.uk/issue/25/app-profiles/.

[16] Gerald Hiebel, Martin Doerr, and Øyvind Eide. “CRMgeo: A Spatiotemporal Extension of CIDOC-CRM”. In: *International Journal on Digital Libraries* 18.4 (Nov. 2017), pp. 271–279.

[17] Fotis Jannidis. “Modeling in the Digital Humanities: a Research Program?” In: *Historical Social Research, Supplement* 31 (2018), pp. 96–100.

[18] Evangelos Kalampokis, Dimitris Zeginis, and Konstantinos Tarabanis. “On modeling linked open statistical data”. In: *Web Semantics: Science, Services and Agents on the World Wide Web* 55 (2019), pp. 56–68.

[19] Werner Kuhn, Tomi Kauppinen, and Krzysztof Janowicz. “Linked Data – A Paradigm Shift for Geographic Information Science”. In: *Geographic Information Science*. Ed. by Matt Duckham et al. Vol. 8728. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 173–186.

[20] Albert Meroño-Peñuela and Ashkan Ashkpour. “Historical Quantitative Reasoning on the Web”. In: *European Social Science History Conference (ESSHC)*. 2016.

[21] Albert Meroño-Peñuela et al. “CEDAR: The Dutch Historical Censuses as Linked Open Data”. In: *Semantic Web* 8.2 (2017), pp. 297–310.

[22] Theodor H. Nelson. “Complex Information Processing: A File Structure for the Complex, the Changing and the Indeterminate”. In: *Proceedings of the 1965 20th National Conference*. ACM ’65. ACM, 1965, pp. 84–100.

[23] Dominic Oldman and Diana Tanase. “Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace”. In: *The Semantic Web – ISWC 2018*. Ed. by Denny Vrandečić et al. Vol. 11137. Lecture Notes in Computer Science. Cham: Springer, 2018, pp. 325–340.

[24] Silvio Peroni and David Shotton. “FaBIO and CiTO: Ontologies for Describing Bibliographic Resources and Citations”. In: *Journal of Web Semantics* 17 (2012), pp. 33–43.

[25] Molly Roberts et al. “Structural Topic Models for Open-Ended Survey Responses”. In: *American Journal of Political Science* 58.4 (2014), pp. 1064–1082.

[26] Muriel Van Ruymbeke et al. “Implementation of multiple interpretation data model concepts in CIDOC CRM and compatible models”. In: *Virtual Archaeology Review* 9.19 (2018), pp. 50–65. URL: https://polipapers.upv.es/index.php/var/article/view/8884.

[27] York Sure-Vetter et al. “Building Knowledge Graphs from Survey Data: A Use Case in the Social Sciences”. In: *The Semantic Web: ESWC 2019 Satellite Events*. Lecture Notes in Computer Science. Springer, 2019, pp. 285–299.

[28] Barbara Tillett. “FRBR and Cataloging for the Future”. In: *Cataloging and Classification Quarterly* 39.3 (1005), pp. 197–205.

[29] Nils B. Weidmann, Jan Ketil Rød, and Lars-Erik Cederman. “Representing Ethnic Groups in Space: A New Dataset”. In: *Journal of Peace Research* 47.4 (2010), pp. 491–499.

[30] Erwin K. Welsch. “Hypertext, Hypermedia, and the Humanities”. In: *Library Trends* 40.4 (1992). Electronic Information for the Humanities, pp. 614–646.

[31] Cord Wiljes et al. “Towards Linked Research Data: An Institutional Approach”. In: *3rd Workshop on Semantic Publishing (SePublica)*. Vol. 994. CEUR Workshop Proceedings, 2019, pp. 27–38. URL: http://ceur-ws.org/Vol-994/paper-03.pdf.

[32] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.160018 (2016).

[33] Benjamin Zapilko and Brigitte Mathiak. “Performing Statistical Methods on Linked Data”. In: *Proceedings of the 2011 International Conference on Dublin Core and Metadata Applications*. DCMI’11. Dublin Core Metadata Initiative, 2011, pp. 116–125.