On the Way to Semantic Interoperability for Historical Data: the Data for History Consortium

EADH2018, Galway, 9 December 2018

1. Introduction: The Semantic Data Challenge in Historical Research

George Bruseker, FORTH-ICS CCI

This panel will introduce the vision and work behind the Data for History (DfH) consortium (dataforhistory.org). DfH is an international consortium founded in 2017 holding the aim of improving geo-historical data interoperability in the semantic web. This aim entails establishing common methods for modelling, curating and managing data in historical research. Such methods would provide foundational support to research projects adopting a framework of collaborative, cumulative and interoperable scientific data production and investigation. DfH aims to develop and then maintain a common ontological model that would allow for domain specific, semantically robust data integration and interoperability. The model(s) will be built up as interoperable extension(s) of CIDOC CRM [1] and relevant ontologies, using the experience of symogih.org [2] and other participating projects, in order to integrate to the wider research community.

In contemporary research, an essential part of a historian's research effort consists in laying the groundwork for scholarly argumentation by investing significant time in the production of complex structured data. This structured data encodes the scholar's work of discernment of unique facts from documents and makes them potentially accessible either as individual facts or in aggregate [3]. Such structured data sources are an invaluable tool of the contemporary historian, allowing for a much more granular and easily repeatable testing of historical argumentation than unstructured data. While nothing can replace the monograph/article in its function as the basic means to deliver complex argumentation, the advent of readily available primary facts encoded in a discrete way gives new means to confirm/disconfirm the arguments proposed therein, by checking arguments against individual and aggregate facts in a digital environment.

To maximize the utility of such structured facts, historians, as in other disciplines, face the task not only of creating such data but of doing so in a standardized way such that data regarding comparable facts and referents are transparently recorded and comparable [4]. The particular challenge for historical research, however, is the necessarily broad horizon of interest of the domain and the consequent significant intellectual challenge in considering how to derive such compatible schema and referent systems. The remit of historical research is so broad that there is not unfounded scepticism at the possibility of coming to any agreement on such questions. Without such agreement, however, the valuable integrated primary data that would allow a more granular investigation of historical argumentation will remain only locally useful, severely limiting its potential use and impact for historical research.

DfH proposes to meet these challenges by adopting the methods of semantic data encoding using formal ontologies. Such ontologies provide a means for historians, working together with computer scientists, to meet this challenge, by co-designing common conceptual models and reference data sources, which would establish a sort of interlingua to exchange data relevant to specific areas of historical research [5]. This panel will offer a means for researchers belonging to DfH to lay out a vision and specific strategies for approaching this question.

[1] CIDOC Conceptual Reference Model (ISO21127:2014): http://www.cidoc-crm.org/
[2] Beretta, Francesco, "L'interopérabilité des données historiques et la question du modèle : l'ontologie du projet SyMoGIH", Brigitte Juanals / Jean-Luc Minel (éds.), Enjeux numériques pour les médiations scientifiques et culturelles du passé, Presses universitaires de Paris Nanterre, 2017 (cf. HalSHS)
[3] Bradley, J. “Silk Purses and Sow's Ears: Can Structured Data deal with Historical Sources?” International Journal of Humanities and Arts Computing. Vol. 8 No. 1, April 2014 (Digital Methods and Tools for Historical Research: A Special Issue): 13-27.
[4] Bruseker, George, Nicola Carboni, and Anaïs Guillem. “Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM.” Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data. Quantitative Methods in the Humanities and Social Sciences. Springer, 2017.
[5] Doerr, Martin, and Nicholas Crofts. 1999. “Electronic esperanto–The Role of the Oo CIDOC Reference Model.” Proceedings of the ICHIM ‘99, Washington DC, 22–26 September 1999.

2. A standard model for a prosopography of religious orders

Bernard Hours, Université de Lyon LARHRA
Georg Vogeler, Universität Graz ZIM-ACDH

The prosopographical approach can renew the understanding of the history of religious orders by introducing the notion of curricula and careers into a historiography nourished either by a sociological approach, or by an analysis of political or economic governance, or by the history of spirituality. Many studies have accumulated more or less structured data, collected in a more or less systematic way and which remain more or less accessible to the scientific community. We are therefore confronted with data heterogeneity and dispersion.

Several projects are already underway to try to remedy this situation. In Austria, the project of digital prosopography of religious orders piloted at the University of Vienna by Professor Thomas Wallnig in connection with the Austrian Centre for Digital Humanities directed by Professor Georg Vogeler [1]. In France within the LARHRA (CNRS Research Center for early modern and modern History), Professor Bernard Hours directs the Monastica project. The latter project will gather data about french Carmelite nuns [2], nuns of several orders who were living dispersed across a large area from North Italia to Spanish and then Austrian Netherlands from 16th to 18th century [3], and finally Cistercians monks in the early modern period [4].

The career of a nun or a monk is broken down into compulsory stages (postulancy, clothing, novitiate, vows) and into the exercise of various formal offices, or fixed period offices, according to variable designation procedures. It takes place within the institution, but it can also take place at the provincial level or at that of the order as a whole. It can therefore lead to greater or lesser geographical mobility. Moreover, in some countries and at certain times, nuns or monks would undergo the experience of exile, an important phenomenon to investigate and understand.

A reflection on the data modelling of this information has already been developed within the symogih.org ontology. The challenge of integrating the project within the DfH consortium is therefore to align the existing model with the CIDOC CRM standard in order to facilitate interoperability between the various projects dealing with the same historiographic problem. The use of CIDOC-CRM offers an opportunity for convergence of the approaches within the DfH Consortium and its interest group "Prosopography use case overview" (B.Hours, G. Vogeler). Indeed, the extension of the CIDOC CRM for the social world, in which the LARHRA is actively involved, will make it possible to have a standard ontology thanks to which data on monastic and religious careers can be aligned.

[1] https://f-origin.hypotheses.org/wp-content/blogs.dir/971/files/2017/01/2017-02-21_CHC.pdf
[2] http://symogih.org/?q=type-of-information-record/82&lang=en: temporal entity (“information”) : « vows »
[3] https://lodocat.hypotheses.org/
[4] http://www.sudoc.abes.fr//DB=2.1/SET=1/TTL=1/SHW?FRST=1

3.APIS: Mapping the Austrian Biographic Dictionary to CIDOC CRM

Matthias Schlögl, Österreichische Akademie der Wissenschaften ACDH

The Austrian Biographical Dictionary (ÖBL) contains well over 18.000 biographies of important people who died between 1815 and 1950. While this resource has a lot of (scientific) shortcomings, it is probably the only dataset that gives an overview of the Austrian-Hungarian Empire and succeeding states (the first and second republic) from the people's perspectives. Other projects have already successfully shown that (automatically) structuring biographical data can give researchers an overview of societal structures that cannot be achieved by non-digital techniques [1]. In 2015 three institutes of the Austrian Academy of Sciences launched the APIS (Austrian Prosopographical Information System) project. The ultimate goal of this project is to semantically enrich the ÖBL and link it to the LOD (Linked Open Data Cloud). To achieve this goal a web application that researchers can use to semantically annotate subsets of the ÖBL by hand has been developed. This Virtual Research Environment (VRE) [2] offers direct access to reference resources – such as "Geonames" [3] or the "Gemeinsame Normdatei" [4] - and makes it an easy task to annotate semantic relations between portrayed persons and other entities (places, persons, institutions etc.). These manual annotations on the other hand can then be used to train machine learning algorithms for automatically extracting entities and relations between these entities [5].

In order to ease the development process of the VRE, the project team decided to use well proven and often used standard web development technologies (Python, Django, MySQL). Similarly we decided to create a very simple and tailor-made data-model, instead of implementing already existing, but more complex ontologies. This decision was based on two rationales. On the one hand we did not want to invest the time needed to develop a VRE that is capable of ontologies, while the project needed only a fraction of the complexity such ontologies offer in order to model the data created. On the other hand, we always planned on creating several versions of our data using various ontologies.

APIS therefore evaluated four different tools/techniques that allow the mappings of existing data to ontologies and the subsequent automatic generation of an RDF data form: D2RQ [6], Karma [7], 3M [8], and Ontop [9]. After critical analysis, our project adopted the Karma tool. In this paper we will present our analysis and focus on a description of the use of Karma for data transformation to an ontology. The advantage of adopting a robust mapping technology like Karma is to allow data generation at a level of complexity that is manageable by the VRE while still being able to re-express the data through mapping technologies even into rich ontology structures such as CIDOC CRM in order to create interoperability with wider data sets. In our contribution we will show the advantages and disadvantages of our workflow (simple data-model/relational database >> mapping tool " complex RDF mapped), discuss the mapping tools and, as an example, run through the mapping process of a subset of our data to CIDOC CRM with Karma.

[1] C. Warren, D. Shore, J. Otis, und L. Wang, „Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks.", DHQ, Bd. 10, Nr. 3, 2016.
[2] M. Schlögl, acdh-oeaw/apis-core: Austrian Prosopographical Information System. 2017.
[3] Geonames: http://www.geonames.org/
[4] http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html
[5] M. Schlögl und K. Lejtovicz, „A Prosopographical Information System (APIS)", in Proceedings of the second Workshop on Biographical Data in a Digital World, 2017. (Forthcoming)
[6] D2RQ: http://d2rq.org/
[7] Karma: http://usc-isi-i2.github.io/karma/
[8] Mapping Memory Manager: http://www.ics.forth.gr/isl/index_main.php?l=e&c=721
[9] Ontop: https://ontop.inf.unibz.it/

4. Best practices for making data generated in automated linkage procedures readily re-usable

Lodewijk Petram, Huygens ING
Jelle van Lottum, Huygens ING
Rutger van Koert, KNAW Humanities Cluster

Since people are central to almost all research in (art) history, and the same persons are often relevant for multiple projects, interoperable and readily re-usable persons data are key to a well-integrated Linked Open Data (LOD) cloud of historical datasets. A common schema for describing person entities would have been invaluable, but the opportunity for proposing such a schema seems to have long passed, given the huge number of person observations already available in online structured datasets and the multitude of models used to describe them. This need not be an insurmountable problem for achieving data interoperability, however, since there is relatively limited variation between the models, allowing for mapping of equivalent classes.

However, there is a category of persons data for which the issues of interoperability and re-usability pose more of a challenge: data that have been generated in automated linkage procedures – a category of data that is set to explode in size in the coming years. Our paper will explore these issues, investigate and assess the ways they are currently being dealt with [1], distil best practices, and reflect on the possibilities for setting a domain standard along the lines of the Data for History consortium, i.e. ensuring semantically robust integration and interoperability of this class of persons data.

In this paper, we take a research project as a use case to explore the issues that we wish to shed light on. The project [2] tries to gain historical insight into the economic contribution of migrant workers on a recipient economy, i.c. the 18th-c Dutch economy. To this end, it reconstructs the careers of maritime workers by applying (semi-)automatic record linkage on the biographical data observations contained in the muster rolls of the Dutch East India Company (almost 800,000 records), to be supplemented in the near future with observations from other sources.

Our algorithms produce suggested matches on the basis of name similarity measures and a set of additional linkage rules for dates and geographical locations. From our research perspective, this linkset contains richer data than the dataset with the original data observations. But, as with all automatic linkage procedures, our method involves a linkage selection bias: e.g. sailors with non-standard names are overrepresented in the mini-iographies we compose. So, for our matched records to be re-usable by other researchers, it is critical for them to know what source data we used, which linkage methods and rules, inter alia, we applied, etc., so that they can decide for themselves whether re-using the records in the linkset is appropriate for answering their questions.

We argue that all relevant linkage information should be made available as provenance data with each record in the linkset and explore ways of properly conveying this information using ontologies (e.g. PROV [3], the P-PLAN [4], extensions to CIDOC-CRM such as CRMsci [5] or the GRaSP model [6]). Finally, we propose best practices for the semantically robust integration of data generated in automated linkage procedures in the LOD cloud.

[1] E.g. Ockeloen, N., Fokkens, A.S., ter Braake, S., Vossen, P., de Boer, V., Schreiber, G., Legêne, S. (2013). Biographynet: Managing provenance at multiple levels and from different perspectives. In Proceedings of the Workshop on Linked Science (LISC2013) at ISWC (2013).
[2] HUMIGEC: https://www.clariah.nl/projecten/research-pilots/humigec
[3] PROV-O: The PROV Ontology: https://www.w3.org/TR/prov-o/
[4] Garijo, D., Gil, Y. (2012). Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data. In Proceedings of the 2nd International Workshop on Linked Science: 12/11/2012, Boston, USA.
[5] CRMsci: http://www.ics.forth.gr/isl/index_main.php?l=e&c=663
[6] van Son, C., Caselli, T., Fokkens, A., Maks, I., Morante, R., Aroyo, L., Vossen, P. (2016) GRaSP: A Multilayered Annotation Scheme for Perspectives. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), pp. 1177-1184.

5. Building a domain specific Research Ontology from external Databases of Academic History.

Thomas Riechert, Hochschule für Technik, Wirtschaft und Kultur Leipzig AKSW
Edgard Marx, Hochschule für Technik, Wirtschaft und Kultur Leipzig AKSW
Jennifer Blanke, Herzog August Bibliothek

The collaborative project: “Early Modern Professorial Career Patterns – Methodological Research on Online Databases of Academic History (PCP-on-Web)” (HTWK Leipzig, HAB Wolfenbüttel) funded by the DFG [1], focuses on domain-specific research ontologies [2]. As is befitting a project that aligns itself with the Digital Humanities, it is interdisciplinary in nature and innovatively combines classic historiographical research methods with Semantic Web technologies in order to investigate scholarly career patterns; a classic prosopographic research question that addresses a significant lacuna in the field that demands to be investigated.

In this talk, we will discuss the results of our current research and will demonstrate how a domain-specific research ontology, which gathers information from several different online databases, is being built. As we will show, PCP-on-Web uses RDF standards for collecting facts (RDF triples) and for describing the vocabulary in a formal way using OWL. Furthermore, the alignment to Data for History (DfH) [3] vocabularies enables the extension of the research ontology through the use of facts from prosopographical databases, as well as the reusability of the resulting research ontology.

The process for building the research ontology is using historical expertise and knowledge engineering methods in parallel. The process covers the database layer, the application layer as well as the research interface layer of the Heloise Common Research Model (HCRM) [4]. Researchers start by exploring available external databases in the way of querying it. These Queries are formally defined by SPARQL [5] and can be used by online access available through SPARQL endpoints or be explored by local tools like KBox [6]. By formal definition, these SPARQL queries represent parts of external databases. They extract relevant concepts and properties for the research vocabulary [7], and can be used for automatic transformation of relevant data into the envisaged research ontology as well. This workflow enables researchers to rebuild the research ontology at any time in the future, so long as the syntax and semantics of the sources are not changing. The usage of a common vocabulary, such as the one to be developed and evolved by the DfH consortium, can mitigate against the problem of inconsistent data. Additionally, the effort of exploring new databases can be minimised, as SPARQL can be defined for a common vocabulary. The usage of external data can be facilitated without resorting to manual exploration.

[1] Research Project: http://pcp-on-web.htwk-leipzig.de
[2] PCP-on-Web research ontology: http://pcp-on-web.htwk-leipzig.de/data/ontology/
[3] Data for History Consortium: http://data-for-history.org
[4] Collaborative Research on Academic History using Linked Open Data: A Proposal for the Heloise Common Research Model, Riechert, Thomas and Beretta, Francesco; In. CIAN-Revista de Historia de las Universidades, 19. (2016)
[5] SPARQL Query Language for RDF: https://www.w3.org/TR/rdf-sparql-query/
[6] KBox: Transparently Shifting Query Execution on Knowledge Graphs to the Edge by Edgard Marx, Ciro Baron, Tommaso Soru, and Sören Auer in 11th IEEE International Conference on Semantic Computing, 2017, San Diego, California, USA.
[7] PCP-on-Web research vocabulary: http://pcp-on-web.htwk-leipzig.de/data/vocabulary/

6. OntoMe : an ontology management environment for extending the CIDOC CRM to historical research sub-domains

Francesco Beretta, CNRS Université de Lyon LARHRA
Djamel Ferhod, CNRS Université de Lyon LARHRA
Vincent Alamercery, Université de Lyon LARHRA

In the domain of historical data interoperability [1] one of the major points under discussion is about the possibility of sharing a common ontology for modelling the whole of human activities in the past. Historians generally think that data produced according to their research agenda in a specific research sub-domain are not reusable in other contexts. To cope with this issue the symogih.org project (Système modulaire de gestion de l'information historique) —started in 2008 with the aim of producing a virtual research environment for collaborative data production— applied a basic distinction between the research agenda of the scholar and the design of a data model conceived as the most objective possible representation of « historical facts ». This allowed the production of a shared vocabulary about specific sub-domains of historical research providing data interoperability among research projects hosted in the platform [2].

This perspective has been broadened since 2013 with the aim of sharing data produced in the symogih.org virtual research environment with other resources available on the semantic web using RDF technologies [3]. For this purpose the project adopted the CIDOC CRM as a conceptual framework which is not only a standardized (ISO 21127:2014) "formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information" but also a generic data model designed with holding the aim to "maintain and support a global knowledge network" [4]. Although this model is generally considered by specialists to be well suited for modelling data in humanities and especially in historical research [5], its high degree of abstraction —an indispensable condition for genericity— poses difficulties when used for data production in historical sub-domains. The solution recommended by the CRM consists in the creation of project-specific extensions but this operation is not easily accessible to the non-specialists.

To support the process of CRM extensions management, and foster the coherence and interoperability of the ontology model development in the domain of historical research, an ontology management environment (OntoME) [6] is currently under development which is designed to facilitate the understanding of the CRM (and of other standardized ontologies and vocabularies) and the production of sub-domain specific extensions submitted to a validation process by the expert community. The platform will allow, on the one side, to import existing data models in the domain of historical research (or even in a wider spectrum) and to map them to the CRM classes and properties with the aim of providing interoperability for project data in the semantic web. On the other side, the platform will support a controlled development process of CRM extensions specific to sub-domains of historical research, allowing to produce explicit sub-classes and sub-properties of the existing, but more abstract ones, and to bundle them into application profiles which can be used for local data production. The paper will present the main components of OntoMe and provide an example of alignment with the CRM of some classes of the symogih.org ontology implemented in the SIPROJURIS project [7].

[1] Meroño-Peñuela Albert, Ashkpour Ashkan, van Erp Marieke, Mandemakers Kees, Breure Leen, Scharnhorst Andrea, Schlobach Stefan, van Harmelen Frank, « Semantic Technologies for Historical Research: A Survey », in Semantic Web – Interoperability, Usability, Applicability (IOS Press) 6(2015): 539-564.
[2] http://symogih.org/?q=documentation
[3] http://symogih.org/?q=rdf-publication – Cf. Beretta Francesco. L'interopérabilité des données historiques et la question du modèle : l'ontologie du projet SyMoGIH. Enjeux numériques pour les médiations scientifiques et culturelles du passé, Paris, Presses Universitaires de Paris Nanterre, 2017, 87-127.
[4] Martin Doerr et al., 'The Dream of a Global Knowledge Network: A New Approach', Journal on Computing and Cultural Heritage, vol. 1, no. 1 (2008).
[5] Courtin, A., Minel, J.-L. (2017). Propositions méthodologiques pour la conception et la réalisation d'entrepôts ancrés dans le Web des données. Enjeux numériques (cit.), 53-86:61-62.
[6] http://ontologies.dataforhistory.org/
[7] http://symogih.org/graph/siprojuris-sym – http://siprojuris.symogih.org/