An ontology for modelling the social, spatial, and semantic relations in pre-modern written sources: Takeaways from data model development in the Dissident Networks Project (DISSINET)

David Zbíral; Adam Mertel; Robert L. J. Shaw; Tomáš Hampejs

Masaryk University, Faculty of Arts, Department for the Study of Religions, Centre for the Digital Research of Religion

The extent of data collection in computational history is often delimited by the specific hypotheses that drive the research in question. Such a parsimonious approach is completely logical and in many cases sufficient; moreover, there is no such thing as "total" data collection, because the data is to a degree in the eye of the beholder. At the same time, however, historical research has a tried and tested tradition of more source-driven research, where the close reading of sources often drives the direction of study more than the testing of hypotheses. In this paper, we present our experience of developing a thorough data model and user interface for the collection of structured data from medieval inquisitorial registers, focusing mainly on the social, spatial, and semantic relations between historical actors, groups, places, physical objects, concepts, and events. We undertook this as part of a project that seeks to provide a networked perspective on religious dissent and its repression in medieval Europe (Dissident Networks Project / DISSINET, https://dissinet.cz). In this paper, we would like to discuss our data model and data collection practices as well as to open the data model to suggestions on how it can be mapped upon existing standards in order to enhance its interoperability.

From our experience, we derive several proposals which could be of interest to historians who, on the continuous scale between hypothesis-driven and source-driven data collection, lean somewhat more towards the latter. Our point of departure is that a data model for source-driven data collection should allow as much relational complexity as the natural language of our sources does. Our approach is not completely new from a conceptual or technical point of view; it is based on statements composed of predicates and actants (subjects and objects), and therefore close to the idea of "semantic triples". However, we dig quite deeply into the language of our sources to propose a way of recording its minutiae, allowing for fuzziness and uncertainty, modifiers (e.g., adjectives, adverbs), temporal and spatial relations (incl. relative chronology), modality (negative, question, possibility, desirability etc.), and give specific meaning to the different actant positions (subject, objects) of each verb for analytical purposes. We thus join the minority strand in current computational history which departs from the idea of factoids, decided upon at the moment of data collection, and rather translate the source into structured data quite extensively to later adapt the specific data projections to the needs of particular research questions and hypotheses we want to test.

This approach allows us to preserve the semantic structure and detail of the sources, while also producing highly structured data suited to various kinds of computational analyses such as social network analysis, socio-semantic network analysis, geospatial data analysis, various regression models, etc. Data collection (or rather, production), in our view, thus already amounts to computational modelling – i.e., in the first instance, we model the source itself, preserving its original language, its vagaries, and its complexity, and only at a later stage various research problems in our focus.

The talk does not focus so much on technical solutions (e.g., review of available data collection environments) or standards. Rather, we explore conceptual issues and a practical workflow that we believe can be inspirational for computational history more generally. Nevertheless, the paper includes a brief demonstration of the *inkVisitor* software, a web-based open-source data collection environment currently under development in our project.