Making better connections in historical research
A knowledge graph ontological extension
considering distortions, vagueness and interpretations

Francis Harvey\textsuperscript{1,2}, Wiesława Duży\textsuperscript{1,3}, Elżbieta Kwiecińska\textsuperscript{1}, and Marta Kuźma\textsuperscript{1}

\textsuperscript{1}University of Warsaw, Faculty of History
\textsuperscript{2}Leibniz Institute of Regional Geography
\textsuperscript{3}Polish Academy of Science, History Institute

1 Introduction

In digital humanities research, the process of connecting source materials readily outstrips the modelling capacities of relational databases usually used today. An ontology modelled as a relational database provides a rigid structure challenged by the multiple connections in interpretative historical analysis. Adaptations to the hermeneutic process of interpretation aid the research process of discovery, analysis and synthesis and can support considerations of distortions in historical sources along with vagueness. The extensions of the historical interpretative research we describe rely on a graph-based data model to support and document a research process that makes better connections than structured database models. The People, Places and Events (PPE) project commenced in late-summer 2021, and we focus on conceptual issues in this extended abstract. The knowledge graph approach in the PPE project builds on ontologies and data from previous projects, OntoHGIS and HOUSe. It uses CIDOC-CRM for a graph-based data model that accounts for knowledge granularity, temporalities and semantic stability of digital representations. With this flexible data structure, research relevant attributes, and interactive interpretation of historical connections, the approach can support historical hermeneutical research more adequately. This approach provides graphic and textual support for semiotic and semantic analysis activities. The data representation and implementation should support close and distant reading techniques, the critical assessment of authorial intentions, multiple perspectives, the constitution of events, and the historical and interpretative process of generating meaning from multiple sources. This approach uses dashboards and can support other interfaces to enrich the digital humanities research process. It can, for example, support systematic approaches to data disambiguation by making both idiomatic and general data and analysis available or source critique to account for historical uncertainties and historiographic biases. It also supports graphic presentations
and geovisualisation that provide other semiotic means to enrich the interpretative research process. Moreover, it also supports the creation of synchronised historical analysis narratives. Currently, only a few months into the project, for the GraphHum presentation in February 2022, we describe in this extended abstract where the project stands in November 2021 and what our next steps will be. To organise the extended abstract, we organise the text into sections considering approaches and issues from previous contributions, the linked property graph approach under development and results from a preliminary pilot study. We also present further research and describe the availability of the extension and its source code for researchers.

2 Previous Work

The PPE project builds on digital humanities research, hermeneutics, knowledge graph concepts and related research projects. CIDOC-CRM specification provides a valuable reference for the design of project data models. PPE also turns to OntoHGIS and HOUSE for some refinements. Given the technical possibilities to translate from XML to Linked Property Graph models, we are keen to assure import and export interfaces. A significant challenge for all systematic ontologies and interoperability is semantics, which technical interoperability specifications and approaches can only partially assure. Due to these concerns, we rely on existing and codified technical interoperability while striving in the project for semantic augmentation. The enhanced semantic capacities will come from the implementation of hermeneutical source critique approaches, which offer an information processing approach and multi-layered support for interactive, and diverse historical interpretation (Liu, 2020). Disambiguation of geographical references (Alberto et al., 2018) and uncertainty (Myrda et al. 2020; Grossner and Meeks 2014; Bodenhamer et al. 2013) matters frame pragmatic foci on assuring interpretational contexts in systematic knowledge approaches. For example, the project 'Entangled ordinances' uses linked open data to support legal history.

For the graph aspects, PPE turns to published work on the use of networks in historical analysis, especially critical reflections that suggest ways to avoid the flight into reductionist objectivism that quickly can emerge from work with digitised data, both large and small data collections. A digital hermeneutics (Fickers, 2020) considers the entire work process from research question formulation, information search, analysis, interpretation, representation and presentation. In an era of abundant historical digital data, the broad concerns consider new digital history practices. In consideration of hermeneutics in the social sciences, Gadamer’s admonition articulates the importance of a focus of interpretation on reflexive engagement with source materials and the process of developing new knowledge (Rodi, 1990). Source criticism has developed to address the ontological differences digital data create, including the fragmentation of the archive and the increasing challenge to define what the original is.

3 Linked Property Graph Approach

The broader issues of digital hermeneutics frame the potential for using Linked Property Graphs to facilitate reflexive engagements. Indeed many of the challenges remain relevant (Weingart, 2018). Data modelling and representation issues can be compared to the differences between structured data representation using the Resource Description Framework (RDF) and Linked Property Graphs (LPG). Triple stores, such as
RDF databases, work with the notion of triples (subject-object-predicate), specified in a specification, usually formalised as an ontology and implemented through data structures based on tuples to store, organise and maintain data sets of arbitrary size. An object-oriented data store keeps data with object IDs (OIDs) following a class hierarchy specified in the ontology. Also associated with the term knowledge graphs, this approach at first glance can seem similar to network approaches. Existing approaches are technologically mature and well-tested but have non-optimised graph data representations and complex queries. In contrast, native graph databases maintain and process graphs, reorganising the data model via the links among elements to optimise the graph data representation. This inherent flexibility of LPG data models is a capability that can support hermeneutical reflexivity in an analytical interpretative process.

A labelled Property Graph is a mathematical organisation stored as data structures, commonly as directed graphs, adjacency matrices, adjacency lists or edge lists with distinct storage and performance strengths and weaknesses (Timón-Reina et al., 2021). A native graph database, such as TigerGraph or Neo4j, is a software system designed to maintain and process graphs storing data as nodes and edges with their attributes through system architectures that rely on capabilities developed for RDBMS and NoSQL databases with REST APIs (Alin et al., 2019). The data organisation uses topological storage management to minimise memory needs and maximise computation with graph elements (for example, in TigerGraph, (Alin et al., 2019)). These backend functions provide a level of database versatility that supports the logical analysis of data relationships in addition to queries known from relational database work.

Overview of development concepts A primary goal for the PPE project is developing an LPG approach that builds on LPG capabilities to provide the abilities to support the concept and semantic changes that emerge in the process of historical interpretation, both in distant and close reading, in the sense of hermeneutical reflexivity. Matters of knowledge granularity, temporalities and semantic stability of digital representation are both informational and knowledge dimensions that arise in work interactions with data and information. Historical data ontologies offer an initial structure that the LPG can flexibly implement to support a range of interpretative (and ultimately also expository) representations. Visualisation plays an important role. It offers visual means of communication for the reflexive hermeneutic process to augment textual forms of communication (Foucault, 1983). A simple expression of the concept is that the graphical elements (Bertin 1985; Bertin 1983) show support the interpretative process rather than occlude meanings with iconic symbolisation. In this sense, Giorgia Lupi’s concept of non-linear storytelling (Lupi, 2012) which focuses on harnessing graphic elements in multiple layers of signification, support the analytical process of interpretation, which puts a clear emphasis on reflection, not just of the presented graphical elements, but of contextual information, digital and non-digital sources, knowledge and of omissions and erasures that can influence, and, of course, even bias, historical interpretations. Steps towards implementation Currently, the network data representation concepts we are working on are prerequisites for the LPG approach and assuring import and export interfaces. Here we can turn to rich work from the digital humanities on the use of network databases to analyse and visualise historical data (Heuvel et al., 2020). The foci have been on integrating various historical actors as nodes in a network who communicate information amongst them. Notes that indicate attributes and objects of communication support both analysis
and visualisation. The results support further analysis and also make the data accessible through visualisations that readily facilitate in-depth questions from an often overwhelming overview of many network relationships (Alvarez and van den Heuvel, 2014).

We find it helpful to think of the LPG approach in distinction to structured network approaches in terms of data representation instead of data models. Previous work on ontologies, linked data and uncertainty of network analyses and source critique (Düring, 2016) is the foundation; the PPE-project sets out to go back to data representation issues and beyond a pre-defined network structure to support flexible and dynamic interpretation at that level and in the process of historical interpretation. In this sense, we also envision that machine-learning techniques developed for network data analysis and bipartite networks approaches (Epstein and Terlinghini, 2015) can be brought in to augment interpretative hermeneutics, opening possibilities for new types of hybrid digital humanities analysis.

4 Next Activities

The presentation reviews relevant approaches from previous contributions, the linked property graph approach under development, and concepts that guide pilot studies using network databases. The work with network databases is currently beginning. While the project is commencing, the activities include workshops and the development of open-source toolkits for interpretation and exposition, which we hope will support further activities and complement existing work. We will be able to share more empirical examples and further developed concepts in February 2022.

References


Düring, M. (2016). How reliable are centrality measures for data collected from fragmentary and heterogeneous historical sources? a case study. In . F. C. In T. Brugh-


