Coding and analysing socio-epistemic networks - an approach to combine modelling and network analysis

Dirk Wintergrün*[1] and Lalli Roberto*

[1]Max Planck Institute for the History of Science – Germany

Abstract

Data modelling intended to support digital or better computational approaches in history faces multiple challenges. Data are normally scarce or at least incomplete, there are only few cases where data are specifically collected for one well defined project and often data has to be brought together from various domains. The challenge is to find a strategy how to foster interinstitutional and interdisciplinary work to collect and analyse data, and to communicate findings based on this data analysis. The pre-condition for this work is the historically adequate modelling of data. But, how can we define criteria to judge what does "adequate" means? We will argue in our talk, that the basis for this process has to be a theorical approach which helps to structure different types of data based on the epistemic goal historians and in particular historians and philosophers of science or – more general – of knowledge want to reach.
Our talk will sketch one possible approach, show how this leads to data models and tools, and finally which results we have reached based on this analysis so far. In joint work at department 1 of the MPIWG we developed a multilayer approach for knowledge organisation and knowledge dynamics based on the approach of historical epistemology. This theory introduces three interconnected layers of knowledge: the semantic layer, describing the structures of knowledge; the material-semiotic network, representing the physical and formal representation of knowledge; and finally the social layer of actors, which are indispensable for the structuring and restructuring of knowledge (Renn et al. 2016; Wintergrün 2019; Renn 2020).

Based on this approach data models compatible to CIDOC-CRM can be developed which encode these three layers (Kräutli and Valleriani 2017; Kräutli et al. 2018; Wintergrün 2019). This leads to semantic graphs typically stored in triple stores. A natural way to analyse this data is to transform these semantic graphs into networks, for a prototypical implementation see (Wintergrün SPARQLgraph). This directly connects data modelling and historical network analysis, the latter becoming an increasingly importing field in quantitative historical research (Düring 2013; Düring et al. 2016).

Networks can be structurally analysed employing a broad range of tools and methods developed in the context of social network analysis, theories of diffusion in medicine and economics, and theories of complex systems in physics and biology. Of particular, help are here theories and methods which help to understand multilayer structures (Lazega and Snijders 2015).

In (Lalli, Howey, and Wintergrün 2019) first results of this approach applied to a case study on the history of general relativity were published. In our talk we will focus on the

*Speaker

procedural parts of this project and discuss the workflow we have implemented to analyse the data. This workflow covers substantial parts of the research data cycle in the humanities, see e.g. (Puhl et al. 2015).

We developed a data model as basis to integrate heterogeneous data which came from our own extensive database of persons (complied in Filemaker) and external sources – in our case WebOfScience and the extensive database of astronomical and astrophysical sources (ADS). This led to an ontology compatible to CIDOC-CRM, revisions of our model were frequently made on the basis of the changing research questions as result of the historical analysis of the data. On the basis of this ontology, all our research data was then stored in a triple store. The dataset currently contains over 800 scholars with more than 3900 relations which were created as results of scholarly analysis of primary and secondary sources by close reading. The scholar relate to more than 450 institutions which are alos part of our graph. These persons are related to approx. 49000 articles.

As a second step, we had to find ways to analyse the data in an interactive way, so that hypotheses on the data could be tested as flexible as possible. The workflow implemented consists of a tool package (Wintergrün n.d.) which can create networks dynamically out of the data stored in the triple store in a format (graphml) readable by traditional tools for analyzing networks (e.g Gephi and Citoscape) and Python-notebooks used as interactive tools to analyse the data, results can be seen in (Lalli, Howey, and Wintergrün 2019). Being interested on the historical development, a focus in these notebooks is to understand the dynamic change of the complex network consisting of the internal relations of the scholars (the social network) connected to their scientific output documented in the articles they published (the semiotic network). The notebooks combine visual and quantitative analysis.

On central challenge in our approach is to find ways to approximate the semantic level, which is the collection of knowledge elements and their relations. There is no direct access to these cognitive elements in the historical data, so we need proxies in the semiotic network that we might hypothesize to be closely related to the underlying semantic structure. The nodes might for instance be words or phrases that represent relevant concepts and the proxy for the connections might be considered their proximity in written texts. Experimentally, we are employing methods from machine learning in a co-operation project with the Berlin Institute for the Foundations of Learning and Data (BI-FOLD) to identify words and phrase in text corpora which are candidates for representations of nodes in the semantic network. The outcomes of this analysis are then semantically modelled and added to the triple store. We will show the first results in the outlook part of out talk.

Düring, Marten. 2013. 'Historical Network Research'. *Historical Network Research* (blog). 7 January 2013. http://historicalnetworkresearch.org/.

Düring, Marten, Ulrich Eumann, Martin Stark, and Linda Keyserlingk, eds. 2016. *Handbuch Historische Netzwerkforschung: Grundlagen und Anwendungen.* Berlin [u.a.]: LIT-Verl.

Kräutli, Florian, and Matteo Valleriani. 2017. 'CorpusTracer: A CIDOC Database for Tracing Knowledge Networks'. *Digital Scholarship in the Humanities.* https://doi.org/10.1093/llc/fqx047.

Kräutli, Florian, Matteo Valleriani, Esther Chen, Christoph Sander, Dirk Wintergrün, and Sabine Bertram. 2018. *Digital Modelling of Knowledge Innovations in Sacrobosco's Sphere: A Practical Application of CIDOC-CRM and Linked Open Data with CorpusTracer.* UNAM. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_2631831_4.

Lalli, Roberto, Riaz Howey, and Dirk Wintergrün. 2019. 'The Dynamics of Collaboration Networks and the History of General Relativity, 1925–1970'. *Scientometrics*, December. https://doi.org/10.1007/s11192-019-03327-1.

Lazega, Emmanuel, and Tom A. B. Snijders. 2015. *Multilevel Network Analysis for the Social Sciences. Theory, Methods and Applications.* Cham: Springer.

Puhl, Johanna, Peter Andorfer, Mareieke Höckendorff, Stefan Schmunk, Juliane Stiller, and Klaus Thoden. 2015. 'Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften'. 2015.

Renn, Jürgen. 2020. *The Evolution of Knowledge. Rethinking Science for the Anthropocene*. Princeton, NJ: Princeton University Press.

Renn, Jürgen, Dirk Wintergrün, Roberto Lalli, Manfred Laubichler, and Matteo Valleriani. 2016. 'Netzwerke als Wissensspeicher'. In *Die Zukunft der Wissensspeicher. Forschen, Sammeln und Vermitteln im 21. Jahrhundert*, edited by Jürgen Mittelstraß and Ulrich Rüdiger, 35–79. Konstanzer Wissenschaftsforum 7. München: UVK Verlagsgesellschaft.

Wintergrün, Dirk. 2019. 'Netzwerkanalysen und semantische Datenmodellierung als heuristische Instrumente für die historische Forschung (Dissertation)'. Dissertation, Erlangen-Nürnberg: Friedrich-Alexander-Universität. https://nbn-resolving.org/urn:nbn:de:bvb:29-opus4-111899.

—. n.d. *Network-Extensions: Managing Multilayer Graphs with Igraph* (version 0.9.5.0). Accessed 27 June 2019a. https://pypi.org/project/network-extensions/.

—. n.d. 'SPARQLGraph'. GitLab. Accessed 10 August 2018b. https://gitlab.gwdg.de/dirk.wintergruen/SPARQL

Keywords: modelling, history of science, network analysis