Data Scopes: towards transparent data research in digital humanities

Motivation

For quite some time, humanities scholars have been using digital tools in addition to their established methodology to try and make sense of large and expanding data sources that cannot be handled with traditional methods alone. The digital methods have computer science aspects that may be combined with but do not readily fit into humanities methodology; an issue which is still too implicit in scholarly debate. This gives rise to a need for methodological consolidation to structure the combination of digital and established humanities methods. In this paper, we propose an approach to such consolidation, that we call *data scopes* (also see Graham, Milligan, Weingart 2016).
In principle, the methodology is relevant for many humanities disciplines, especially those that deal with large-scale heterogeneous data sets and sources. We think that digital methods should extend, not replace established methodology. Digital tools are now often employed in a methodological vacuum as if they would yield results all by themselves, but we propose that they should always be embedded in broader research methods.

Digital Data and the Research Process

Data and data sets are often presented as external to research, as the data are the sources upon which the research draws. However, all research considers datasets from the vantage point of research questions. Data and questions shape and transform each other in cycles of searching, selecting, close and distant reading, and extending the data with other data sets and annotations. In this process, the scope of the data and the scope of the research questions are aligned so the latter can be addressed. Preparing data for analysis requires interpretation and is therefore inseparably part of research and should be incorporated into the disciplinary methodology. This calls for an extension of usual source criticism with more specifically digital source criticism. In a typical research project, involving digital data, they are processed with a variety of tools that change them in many ways, making tool results and data at times inseparable. Tool and data criticism are therefore intertwined.

We will illustrate our argument with an example from research on the change in discourse regarding migration in Western Europe from 1913-2013 (van Faassen and Hoekstra, 2017). That study focuses on a ‘scientization’ of the migration debate and how the scientists and politicians in the debate were connected to each other. The overall research question can be addressed in many ways, but not straightforwardly answered from a single data set as the discourse spans a very long period and a lot of different media. Therefore the question was split into several specific questions that can be more directly operationalized as analyses of a combination of two digital sources, the *Publications of the Research Group for European Migration Problems* (REMP) and the online *International Migration* bulletins of the Intergovernmental Committee for European Migration (ICEM), which merged with the

REMP-bulletin. For the first dataset, it was possible to identify key actors and their roles and to address specific questions: “who were writing forewords, prefaces or introductions to each other’s work; Who ordered the research? Who financed it? etc.” (van Faassen and Hoekstra, 2017, p. 7). This requires modelling of actors (persons and organisations), their roles and the relationships between them, normalizing names of persons and mapping changing roles and names of roles over time, and linking them across publications. This in turn requires interpretations relying on domain knowledge that need to be argued for. For the long term trends in the migration discourse, the frequency in the occurrence of key terms was analysed using the other dataset (and a control set) consisting of series of article titles in *International Migration* and *International Migration Review*, two important long running journals supplemented by topic overviews from WorldCat. This required not only the use of weighted frequency analysis, but also a critical assessment of the value of the various series. Consequently, preparing data sets and analyzing them tends to happen in iterations, where initial analyses inform a next iteration of selecting, modelling, normalizing and linking, and data and research scopes are brought ever closer together.

Conceptual model

Researchers start a research project with a research question, that may be adjusted and expanded in the course of the research process. From the onset, these questions determine the research scope and therefore the scope of which data are relevant. As the research proceeds, partial questions are either answered, or prove to be unanswerable with the available data, because of their form or because of the nature and extent of the data. Researchers then interact with their data, to annotate them in such a way that it enables them to answer questions. They may also pull in other data sets to expand the existing cluster of data, so that the scope of the data will fit the research scope better. In light of these dynamics of the research process, data and data clusters are not just ‘raw material’ that is object of study, but points of departure that these iterative research interactions change (Boonstra et al. 2006, 23). It is the research process that turns a data cluster into a data scope:

We may discern a limited number of separate activities working that are part of this research process that produces the data scope:

- *Modelling* represents the data in such a way that it will fit the research scope
- *Normalizing* structures data and reduces data variation so that they may be queried more easily and they can be used for comparisons, classifications or calculations
- *Linking* data connects previously unconnected data, providing them with context from other data sets. Researchers should be aware that the validity or relevance of links can be context-dependent (person X and Y are linked for a specific question because they played the same role, but the link may be invalid for other questions, see Brenninkmeijer et al 2012).
- *Classifying* groups data in order to reduce complexity. This adds a level of abstraction to the data

As the research process transforms a cluster of data into a data scope, many levels of annotations are added to the original data sets. For the purpose of our analysis these annotations comprise all types of data enrichment, that range from structuring (for instance by adding markup to a text), to identifying named entities and keywords as structured metadata to adding explanatory notes and everything in between. It is easy to lose track of the changes, as enriching and transforming often goes in small steps and using many different manual and automatic procedures, and because transformations are often cumulative. In light of the incremental transformative effect of the research process upon the data, it is important to keep track of these changes, so that researchers can communicate about and account for their data scopes. Documenting the data changes makes both the research process and the resulting data scope transparent. This way, the research process also becomes reproducible and transferable to other research data clusters. (Groth et al. 2012; Ockeloen et al. 2013)

Discussion

The data scope concept is not a strictly theoretical model, but it is rooted in an experience of many years of empirical research with a lot of different research projects. In the presentation, we will illustrate the concept with an example of a research project that combines a number of data sources to analyze the long-term perspectives on discursive cycles relating to migration and migration policy.

Data scopes are not just a plea to work interdisciplinary and collaboratively, but in a number of research processes they are necessary. While our examples are mostly drawn from the historical sciences, the value of the concept is by no means confined to that disciplinary field.

In our view, such a methodological approach is always indispensable when there are large amounts of data available for research and when the data are of a more heterogeneous nature. Transparency and transferability is also important when researchers from different disciplines collaborate or when there is a collaboration between humanities researchers and more technologically oriented partners. In those cases, researchers, who have often mostly separate tasks, have to make sure that they understand each other to prevent misunderstandings and waste of resources.

Many humanities researchers have already adopted a part of the methodology, but unfortunately, they often just use a number of tools, without acknowledging the cumulative transformative effects these have. The value of the proposed model is in the emphasis on a coherent methodological approach to doing research with (large scale) data.

References

Boonstra, Onno, Leen Breure en Peter Doorn, 2006. *Past, present and future of historical information science,* Amsterdam 2006

Brenninkmeijer, C., et al. 2012. Scientific Lenses over Linked Data: An approach to support task specific views of the data. A vision. http://linkedscience.org/wp-content/uploads/2012/05/lisc2012_submission_8.pdf van Faassen, Marijke, Rik Hoekstra. 2017. *Modelling Society through Migration Management. Exploring the role of (Dutch) experts in 20th century international migration policy*. Conference paper. Government by Expertise: Technocrats and Technocracy in Western Europe, 1914-1973. Panel 3. Global Expertise.

Graham, S., I. Milligan, and S. Weingart. 2016. *The Historian's Macroscope: Big Digital History* http://www.themacroscope.org/2.0/

Groth, P., Y. Gil, J. Cheney, and S. Miles. 2012. "Requirements for provenance on the web." *International Journal of Digital Curation* 7(1).

Ockeloen, N., A. Fokkens, S. ter Braake, P. Vossen, V. de Boer, G. Schreiber and S. Legêne. 2013. BiographyNet: Managing Provenance at multiple levels and from different perspectives. In: *Proceedings of the Workshop on Linked Science (LiSC) at ISWC 2013*, Sydney, Australia, October 2013. http://linkedscience.org/wp-content/uploads/2013/04/paper7.pdf