Documenting Research Practice in DH

Marijn Koolen - KNAW Humanities Cluster
Rik Hoekstra - KNAW Humanities Cluster
Andreas Fickers - University of Luxembourg
Antske Fokkens - VU University Amsterdam
Jasmijn van Gorp - Utrecht University
Jacco van Ossenbruggen - Centrum Wiskunde & Informatica, VU University Amsterdam

Background and Motivation

Digital humanities research is often a complex process involving many data interactions and transformations, leading towards insights and interpretations that are used to construct arguments and narratives. But scholarly communication in the humanities tends to leave out much of this process and focuses on the narrative, with references to sources and software that were used in the analysis. There is a growing concern in DH communities that research processes are not described in sufficient detail to assess how these processes change the nature of the data and how they affect interpretations and their use in argumentation.

There is a lack of shared methodological understanding of how assumptions built into digitization processes and digital tools for gathering, selecting, cleaning and analysis align with or are taken into consideration by the scholars who use them or rely on them.

There are a number of initiatives to develop methodology for dealing with digital materials and methods in the humanities. Digital Source Criticism (Fickers 2013) deals with evaluating a digital source document or object, its creation, distribution and use. Digital Tool Criticism (Koolen et al. 2018) deals with evaluating the use of digital tools in research and understanding its role in the research process. Data Scopes (Hoekstra and Koolen 2018) deals with evaluating the transformations of digital data that are needed to address research questions with them.

Developments of new DH methods have challenges and consequences at three levels:
1. the epistemological aspects of reaching a shared understanding within research communities, such that peers can critically assess and appreciate the data interactions in each other's work,
2. documenting the research process in such a way that a scholar can critically reflect on this process, communicate about it transparently and where possible, share it for purposes of reflection, reuse and reproducibility, and
3. the incorporation of the documentation of the research process into the narrative of scholarly publications.

In this workshop we focus on the level of documentation, as we argue it is needed to deal with the challenges of the other two levels. The code underlying these tools contributes to interpretation (van Zundert 2018) whether the scholar using them is aware of this contribution or not. Documenting the research process, the tools and parameters used (also for the failed experiments and dead ends) allows us to discuss these contributions and their consequences for epistemology.

The workshop will address a number of questions:
1. What are ways of documenting? E.g. keeping research diaries, making screenshots and screencasts, putting datasets and analysis code in online repositories, keeping automatically generated transactions logs of tools (containing e.g. queries, clicks, selections and filters).
    a. Balance between automated (disadvantage: no reflection) and manual documentation.
    b. Move from tools that support interactive data transformations towards interactive tools that build explicit strategies to document and execute such transformations repeatedly (Van Ossenbruggen et al. 2018, Cornacchia et al. 2017, OpenRefine.org)
    c. Jupyter Notebooks - example of unfolding non-linear process to linear process (Burton and van Zundert, under submission)
2. What is the purpose of documenting research?
    a. The individual interactions with tools and data are useful for reuse, reproducibility, and pedagogy, but less for presentation or argumentation and validation/reviewing.
    b. The interpretations, considerations and choices made are useful for argumentation and validation.
3. What are current practices within DH and in other fields for documenting their research? What is documented and what is not?
    a. Look at what digital ethnographers do, e.g. reflect on own research process and its influence on interpretation.
    b. Look at our own projects: reaching a dead end, backtracking and insights gained at those points.
    c. Do we need to develop better vocabularies to document our research process at the right level of abstraction? E.g. when performing commonly used statistical tests, it typically suffices to refer to the name of the test used, without the need to explain the details of the underlying software that performed that test. We lack such commonly used procedures for many other data transformations, which forces us to document these transformations on an often boring and overly detailed level.

Workshop Format

This full day workshop will combine theoretical discussion and experimental sessions in which participants and organisers try out and discuss ways of documenting.

Session 1: The workshop starts with a general introduction to the workshop and its aims, followed by presentations from a few invited speakers from different disciplines, e.g. ethnography, qualitative social science, human computer interaction and digital humanities, who will describe their documentation practices.

Session 2: In the first experimental part, the workshop participants will work in small groups to examine and try to document previous research. The workshop organisers will pre-select and prepare materials for 1 or 2 published data-driven studies in which they were involved, so that they can provide missing information where possible.

Session 3: To wrap up the experimental part, groups will report their experiences and findings, which will be used as a basis for a collaborative publication. The workshop will conclude with a closing discussion about lessons learned and future directions.

References

Cornacchia, R., Hildebrand, M., de Vries, A.P., Dorssers, F. 2017. Challenges for industrial-strength Information Retrieval on Databases. 1st International Workshop on Keyword-based Access and Ranking at Scale. co-located with 20th EDBT/ICDT 2017 Joint Conference 21 March 2017, Venice, Italy. http://ceur-ws.org/Vol-1810/KARS_paper_04.pdf

Fickers, A., 2012. Towards a new digital historicism? Doing history in the age of abundance. *VIEW Journal of European Television History and Culture*, *1*(1), pp.19-26.

Hoekstra, R. and Koolen, M., 2018. Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pp.1-16.

Koolen, M., van Gorp, J. and van Ossenbruggen, J., 2018. Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*.

Van Ossenbruggen, J., Wielemaker, J., Hildebrand, M. 2018. Amalgame: a SKOS vocabulary alignment tool. https://doi.org/10.5281/zenodo.1228926

van Zundert, J.J. 2018. Why We Should Think About a Domain Specific Computer Language (DSL) for Scholarship. Paper presented at the conference "Digital Hermeneutics in History: Theory and Practice", 25-26 October, Luxembourg.