Reflections on reading history from a distance

Max Kemman*, University of Luxembourg, max.kemman@uni.lu
Mark Hill*, London School of Economics, m.j.hill@lse.ac.uk
John Regan*, University of Cambridge, jjr35@cam.ac.uk
Paul Nulty*, University of Cambridge, pgn26@cam.ac.uk
Peter de Bolla, University of Cambridge, pld20@cam.ac.uk
Pim Huijnen*, Utrecht University, p.huijnen@uu.nl
Tom Kenter, University of Amsterdam, tom.kenter@uva.nl
Daniele Guido*, University of Luxembourg, daniele.guido@uni.lu
*Presenting authors

Panel abstract

This panel aims to discuss how digital tools fit in to historical practices, and reflect on the interaction between digital and historical methods. While methods of computational textual analysis have been used in many other disciplines (Laver, Beniot, and Garry 2003, Grimmer and Stewart 2013, Schonhardt-Bailey 2006, Moretti 2013), historians (in particular, intellectual historians) have been suspicious of 'distant reading.' As Skinner argued, historical texts are not positive facts from which we can reconstruct empirical meaning (Skinner 2002). Nonetheless, there is an interest in these ideas and their application to historical studies. Thus, as a case study of the interaction between digital and historical methods, this panel will focus on how tools, and specifically distant reading techniques, may be used to investigate "concepts" in historical documents.

This panel follows recent investigations by the likes of Guldi and Armitage, who argue in *The History Manifesto* that, with the emergence of new digital materials, methods, and techniques, historians may be able to return to *longue durée* historical investigations (Guldi and Armitage 2014). However, how this is exactly to be done is not a triviality, and while most reflections have foregrounded the technical at the expense of the methodological, this panel hopes to investigate this relationship in more detail. That is, there is a shared concern that some digital investigations may introduce unexpected methodological problems and thus may ultimately have an impact on the accuracy of analysis and claimed conclusion. In essence, distant reading depends on counting occurrences of terms, but what these terms mean conceptually may change over time and context. When treating historical sources as big data and interpreting the outcome of distant reading methods, historians risk remaining ignorant to such *concept drift*. These issues have not been ignored (de Bolla 2013, Edelstein 2015, London 2016, Wang, Schlobach, and Klein 2011), but there remains a need for further serious methodological reflection, in addition to technological solutions, if we are to practice this sort of digital history.

To address these issues this panel brings together a number of researchers who are engaging with these problems from differing perspectives. This includes researchers who are at the cutting edge of the use of quantitative text analysis on historical documents, with John Regan and Peter de Bolla aiming to show how historically-sensitive distributional text analysis can facilitate the recognition of conceptual architectures, and Pim Huijnen and Tom Kenter focusing on continuities and changes in conceptual structures over time. Mark Hill looks at ways one may sidestep the traditional methodological problems quantitative analysis may be introducing by turning to niche historical investigations. Daniele Guido will reflect on his experience with both developing tools to be used by historians, and the potential pitfalls which may lurk. Finally, Max Kemman will reflect on how historians collaborate in digital history projects. The goal, then, is to bring together these practitioners, contrast their differing technological and historiographical approaches, and ultimately offer some further reflections on how digital history facilitates an interaction between digital technology and historical practices.

John Regan, Paul Nulty, & Peter de Bolla - What distributional concept analysis tells us about the philosophical concept of 'negative liberty': A case study in the shadow of Quentin Skinner

In his 1984 essay 'The idea of negative liberty', Quentin Skinner gives a historical account of two opposing ideas. One is 'negative liberty', in which the individual's social freedom is guaranteed only by the absence of limiting factors such as state intervention, responsibilities to one's communities, and other externalities. In this scheme, liberty can only be defined negatively, as Thomas Hobbes has it at the start of his chapter 'Of the liberty of subjects' from *Leviathan*: 'liberty or freedom signifieth (properly) the absence of opposition.' Skinner contrasts this with an ideal of liberty in which the operative factor is the virtue and value of public service. This is to say, that one is only consummately free when one acknowledges one's social responsibilities and when one carries out virtuous acts of public service. These contrasting ideas of liberty are named by Canadian philosopher Charles Taylor as the 'opportunity concept' and the 'positive exercise concept'. The former relies purely on the absence of constraint and prescribed social objectives, where in the latter the individual attains liberty by acting positively in the service of the state or community.

This presentation will include data from the dataset *Early English Books Online* (EEBO) and *Eighteenth Collections Online* (ECCO) in order trace the emergence, stability and evolution of the concept of liberty from the seventeenth century to the end of the eighteenth. The software developed by the Cambridge Concept Lab is capable of interrogating these datasets in sophisticated computational and statistical ways, massively enhancing the capabilities of these data resources. We shall share these data and demonstrate the innovative power of our methodology for understanding the history of philosophy.

Pim Huijnen & Tom Kenter - What we talk about when we talk about concepts. Applying distributional semantics on Dutch historical newspapers to trace conceptual change

Word embeddings – vector representations of words that embed words in a so-called semantic space where the vectors of semantically similar words lie close together – are increasingly used for semantic searches in large text corpora. Word vector distances can be used to build semantic networks of words. This closely resembles the notion of semantic fields that humanities scholars are familiar with.

We have previously shown how word embeddings, as produced by a popular implementation word2vec, can be used to trace concepts through time without the dependency of particular keywords (Kenter et al. 2015). However, there are two main challenges that come with the use of word embeddings to represent concepts and conceptual change for the study of history. Firstly: commensurability. The use of computational techniques like word2vec demands choices of practical or technical nature. How do we legitimize these choices in terms of conceptual theory? Secondly: dependency on data. Do the results of word embedding techniques provide insights into real conceptual change, or do they merely reflect arbitrary biases in the underlying data?

Both challenges illustrate the need for critical reflection now that advanced computational tools are adopted in historical scholarship. Based on concrete examples, we will show how we dealt with these challenges in our research.

Mark Hill - Niche Analysis: Historical Methods, Digital Humanities, and Smaller Data

This paper aims to question the relationship between intellectual history – in particular, the

Cambridge School (Skinner) – and new tools and techniques in quantitative text analysis. Specifically, it asks whether one can extract contextually relevant and historically interesting information from a digital corpus via methods of distant reading used on specific historical contexts. That is to say, while the paper does not claim that others have not engaged with these issues (de Bolla 2013), it does aim to offer thoughts on an engagement from a different angle: to investigate the use of these methods in niche historical areas, places, and time.

This approach may provide a number of benefits. First, by limiting itself historically and geographically the project mitigates the problems of decontextualized analysis. Second, by directing these tools towards a topic which one already has an understanding of, the problems associated with 'distant reading' become less pronounced – a researcher can more easily verify or falsify algorithmic outputs (Betti and van den Berg 2014). Third, by working with a limited dataset one is able to construct a more accurate corpus (Bullard 2013, Spedding 2011). Forth, it may allow us to take sources which have been read (and re-read) for centuries and extract new information. Finally, as the project's scale is limited its methodological lessons are more easily replicable, and therefore of use to other scholars.

While the potential outputs are certainly limited when compared to some other projects, the author claims that there may nonetheless be a role for the analysis of smaller sets of data.

Daniele Guido - When It Fails

There are many accessible, low cost or free to use web tools and services that enable humanities researchers to engage with complex analytical algorithms. Backed by the powerful Dbpedia ontology, identifying people, locations or institutions in different media has never been more easy. Tools like DBpedia Spotlight (Mendes et al. 2011), YAGO/AIDA (Hoffart et al. 2011), Textrazor (Van Erp, Rizzo, and Troncy 2013) or Babelfy (Moro, Cecconi, and Navigli 2014) allow researchers to effectively recognize named entities in different languages. Moreover, if those entities are somehow present in DBpedia, these tools can resolve ambiguity relying to DBpedia internal linking. This analysis activity has been made available for images as well, at least for people and groups recognition - with Image Feature Extraction (IEF) (Fang et al. 2015). Both categories of services have been tested and integrated into Histograph (histograph.eu), a web platform for multimedia collection exploration conceived with historians. Histograph enables automatic enriching and annotating of texts and images for different services and for different languages. However, when exploring the results of entity recognition activities, the multitude of homonyms, misspellings and ambiguities makes it clear that alongside automatic tools there is a need for side processes that help us correct the results and reduce noise. This activity of spotting and correcting errors is time consuming work for the researcher, but having unexpected and usually "wrong" results makes him/her start to disbelieve in the expression "the more the information, the better the result", resulting in *diminishing trust* of the tool. In Histograph we adopted a mixed approach, combining a technological solution with design.

On one side, we disambiguate the results by comparing between sources in different languages, obtaining different perspectives and contexts. On the other side, interactive visualizations - which elsewhere have been shown to help digital humanities scholars to evaluate and interpret complex datasets (Jänicke et al. 2016), enable exploration and the identification of *patterns of failures*. By thus combining technology and design we hope researchers accept the failure of the tool whilst they simultaneously "love the bomb".

Max Kemman - Digital History Projects as Boundary Objects

Digital history as a subfield of the digital humanities constitutes a form of *methodological interdisciplinarity*; using methods, concepts, or tools from other disciplines to try to improve historical research (Klein 2014). However, as this panel demonstrates, this is not a straightforward process of

taking something from another discipline and implementing it in historical research. Instead, what we see is a negotiation of practices to align the new methods with the scholarly values of the discipline (Kaltenbrunner 2015). This negotiation regularly takes place in the context of a research project, where participants with different backgrounds work together on a shared problem. Yet despite working on a shared problem, the individual participants may still have different research goals and incentives to enter the collaboration. Although the research project defines a common research problem, how this research problem is or should be approached differs between the different collaborators dependent of, among other factors, their disciplinary background. This paper will therefore analyze the research project as *boundary object*, i.e., as an object that maintains a common identity among the different participants, yet is shaped individually according to disciplinary needs (Star and Griesemer 1989, Star 2010). In order to investigate how participants shape the research project and align the project with their scholarly values, we will look at the individual *incentives* for collaboration, following research by Weedman on incentives for collaborations between earth scientists and computer scientists (Weedman 1998). For several digital history projects, we will discuss collaborators' reasons for joining the project, their individual goals with the project, and the expected effects of the participation after the project has ended.

This research is part of a PhD research on how the interdisciplinary interactions in digital history have methodological and epistemological consequences for the practice of historians (Kemman 2016). By untangling the individual interests in digital history projects, we aim to gain better insight into how digital history functions as a coordination of practices between historians and collaborators from other disciplinary backgrounds.

Bibliographic References

Betti, Arianna, and Hein van den Berg. 2014. "Modelling the History of Ideas." British Journal for the History of Philosophy 22 (4). Informa UK Limited: 812–35. doi:10.1080/09608788.2014.949217.

de Bolla, Peter. 2013. The Architecture of Concepts. Fordham University Press. doi:10.5422/fordham/9780823254385.001.0001.

Bullard, Paddy. 2013. "Digital Humanities and Electronic Resources in the Long Eighteenth Century." Literature Compass 10 (10). Wiley-Blackwell: 748–60. doi:10.1111/lic3.12085.

Edelstein, Dan. 2015. "Intellectual History And Digital Humanities." Modern Intellectual History 13 (01). Cambridge University Press (CUP): 237–46. doi:10.1017/s1479244314000833.

Van Erp, Marieke, Giuseppe Rizzo, and Raphaël Troncy. 2013. "Learning with the Web: Spotting Named Entities on the Intersection of NERD and Machine Learning.." In # MSM, 27–30.

Fang, Hao, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao et al. 2015. "From captions to visual concepts and back." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1473-1482.

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Political Analysis 21 (3). Cambridge University Press (CUP): 267–97. doi:10.1093/pan/mps028.

Guldi, Jo, and David Armitage. 2014. The History Manifesto. Cambridge University Press (CUP). doi:10.1017/9781139923880.

Hoffart, Johannes, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. 2011. "YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages." In WWW.

Jänicke, S, G Franzini, MF Cheema, and G Scheuermann. 2016. "Visual Text Analysis in Digital Humanities." In Computer Graphics Forum. Wiley Online Library.

Kaltenbrunner, Wolfgang. 2015. "Reflexive Inertia: Reinventing Scholarship through Digital Practices." PhD thesis, Leiden University.

Kemman, Max. 2016. "Dimensions of Digital History Collaborations." In DHBenelux. Belval, Luxembourg.

Kenter, Tom, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 2015. “Ad Hoc Monitoring of Vocabulary Shifts over Time.” Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 1191–1200.
Klein, Julie Thompson. 2014. Interdisciplining Digital Humanities: Boundary Work in an Emerging Field. Online. University of Michigan Press. doi:10.3998/dh.12869322.0001.001.
Laver, Michael, Kenneth Benoit, and John Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” American Political Science Review 97 (02). Cambridge University Press (CUP). doi:10.1017/s0003055403000698.
London, Jennifer A. 2016. “Re-Imagining the Cambridge School in the Age of Digital Humanities.” Annual Review of Political Science 19 (1). Annual Reviews: 351–73. doi:10.1146/annurev-polisci-061513-115924.
Mendes, Pablo N, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. “DBpedia Spotlight: Shedding Light on the Web of Documents.” In Proceedings of the 7th International Conference on Semantic Systems, 1–8. ACM.
Moretti, Franco. 2013. Distant Reading. Verso Books.
Moro, Andrea, Francesco Cecconi, and Roberto Navigli. 2014. “Multilingual Word Sense Disambiguation and Entity Linking for Everybody.” In Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272, 25–28. CEUR-WS.org.
Schonhardt-Bailey, Cheryl. 2006. From the Corn Laws to Free Trade: Interests, Ideas, and Institutions in Historical Perspective. Mit Press.
Skinner, Quentin. 1984. “The Idea of Negative Liberty: Philosophical and Historical Perspectives.” Philosophy in History. Cambridge University Press Cambridge, 193–221.
Skinner, Quentin. 2002. Visions of Politics. Cambridge University Press (CUP). doi:10.1017/cbo9780511613784.
Spedding, Patrick. 2011. “The New Machine: Discovering the Limits of ECCO.” Eighteenth-Century Studies 44 (4). Johns Hopkins University Press: 437–53. doi:10.1353/ecs.2011.0030.
Wang, Shenghui, Stefan Schlobach, and Michel Klein. 2011. “Concept Drift and How to Identify It.” Web Semantics: Science, Services and Agents on the World Wide Web 9 (3): 247–65. doi:10.1016/j.websem.2011.05.003.
Skinner, Quentin. n.d. “Meaning and Understanding in the History of Ideas.” In Visions of Politics, 57–
89. Cambridge University Press (CUP). doi:10.1017/ccol0521581052.004.
Star, S.L., and J. R. Griesemer. 1989. “Institutional Ecology, ‘Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39.” Social Studies of Science 19 (3): 387–420. doi:10.1177/030631289019003001.
Star, S.L. 2010. “This Is Not a Boundary Object: Reflections on the Origin of a Concept.” Science, Technology & Human Values 35 (5): 601–17. doi:10.1177/0162243910377624.
Weedman, Judith. 1998. “The Structure of Incentive: Design and Client Roles in Application-Oriented Research.” Science, Technology & Human Values 23 (3): 315–45. doi:10.1177/016224399802300303.