The Ideal Corpus. Towards a Critique of Large Digital Libraries from a Digital Humanities Perspective

Steven Claeyssens
Koninklijke Bibliotheek, National library of the Netherlands
steven.claeyssens@kb.nl

Large scale digitisation of historical paper publications enables analyses of vast amounts of digital surrogates using machines, algorithms and software to ‘read’ the texts. However, continuously expanding collections of such texts, like the Google Books corpora, HathiTrust or – closer to home – Delpher, combined with a proliferation of computational approaches to study textual data and the growing number of scholarly disciplines taking part in the Digital Turn, calls for a renewed reflection on two of the key aspects of the research process: source selection and source criticism.

A recent special issue of the *Tijdschrift voor tijdschriftstudies* on ‘Historical Research and Delpher’ (2015) brings together a number of researchers from a range of disciplines. Based on their experiences with Delpher and the Delpher collections the authors point out various shortcomings and formulate additional requirements, both concerning the data and the search engine. Since the Delpher user interface is just one of many potential tools to mine the Delpher data, this paper focusses in particular on questions relating to the selection of content for Delpher (what, how, when, etc.). An analysis of these types of data-wish lists not only leads to the identification of valuable recommendations for the Delpher team, but also to some incompatible conclusions indicating the conflicting expectations and interests of different disciplines and researchers and the countless number of tools they are using to analyse textual data.

This paper argues that a digital source criticism is urgently needed to tackle the questions raised by these opposing expectations. Researchers and librarians should collaborate closely on this and join forces to define the limits and fits of ‘the ideal corpus’. Inspiration for this definition can, amongst other things, be found in textual criticism, corpus linguistics and analytical bibliography.

References

Broersma, M. (2012) ‘Nooit meer bladeren? Digitale krantenarchieven als bron’, *Tijdschrift voor mediageschiedenis*, vol. 14, no. 2, pp. 29-55.

Dahlström, M. (2010) ‘Critical Editing and Critical Digitisation’, Peursen, W. van, Thoutenhoofd, E.D. and Weel, A. van der (ed.), *Text Comparison and Digital Creativity. The Production of Presence and Meaning in Digital Text Scholarship.* Leiden: Brill.

Fickers, A. (2012) ‘Towards a new Digital Historicism? Doing History in the Age of Abundance’, *Journal of European History and Culture*, vol. 1, no. 1, pp. 19-26.

Flanders, J. and Matthew L. J. (2013) *A Matter of Scale. Keynote Lecture from the Boston Area Days of Digital Humanities Conference*, Boston: Northeastern University.

Nicholson, B. (2013) ‘The Digital Turn. Exploring the methodological possibilities of digital newspaper archives’, *Media History*, vol. 19, no. 1, pp. 59-73.

Pechenick, E.A., Danforth, C.M. and Dodds, P.S. (2015) ‘Characterizing the Google Books Corpus. Strong Limits to Interferences of Socio-Cultural and Linguistic Evolution’, *PLoS One*, vol. 10, no. 10, October, pp. 1-24.

Padilla, Th. (2016) ‘Humanities Data in the Library. Integrity, Form, Access’, *D-Lib Magazine*, vol. 22, no. 3/4, March/April.

Piper, A. (2012) *Book was there. Reading in Electronic Times*. Chicago: University of Chicago Press.

Ramsay, S. (2011) *Reading Machines. Toward an Algorithmic Criticism*. Illinois: University of Illinois Press.

Sijs, N. van der (2014) *De voortzetting van de historische taalkunde met andere middelen. Inaugurele rede*. Nijmegen: Radboud Universiteit.

Tanselle, G.T. (1980) ‘The Concept of the Ideal Copy’, *Studies in Bibliography*, vol. 33, p. 18-53.

Weel, A. van der (2011) *Changing our Textual Minds. Towards a Digital Order of Knowledge*. Manchester: Manchester University Press.