The exploration and visualisation of big data in the humanities,
Comparing data representation of a large-scale e-infrastructure (CLARIN) and a dedicated Virtual Research Environment (Soundbites)

Douwe Zeldenrust
Meertens Instituut (Royal Netherlands Academy of Arts and Sciences)

Introduction

In the past decades the numbers of crossovers between academic research and computing have increased drastically. Efforts to introduce computational methods typically involve collaborative work between scholars and engineers plus the combination of their complementary skills and expertise. Such collaborations can be considered as encounters between 'epistemic cultures' (Knorr-Cetina and Berry). These encounters are combinations of meanings, material arrangements and practices that organize an area of scholarly work. This paper focussus on a particular encounter at a specific time: the current interaction between the new epistemology of 'big data' and the humanities. The present phase of this encounter is marked by a major challenge: the exploration and visualisation of big data.

Big data and the humanities

Since the late 20th century, huge databases have become a ubiquitous feature of science. Big data has become the term for describing this new and distinctive mode of knowledge production. Within the humanities these modern data practices are increasingly visible. In the domain of spatial humanities there also have been practical experiments to access data and metadata in an innovative way. At the 2012 ESSHC, reports were given about several of these experiments. 'Deep mapping', inventive geo-visualisations and the creation of geographic virtual research environments can in this respect be regarded as stepping-stones for exploring and visualising quantitative and qualitative big datasets (Bodenhamer, Gregory and Zeldenrust).

Accessing big data: CLARIN

Since the middle of last the decade investments in large-scale e-infrastructures for the humanities have risen enormously. Projects such as CLARIN, DARIAH and more recently CLARIAH, received funding. CLARIN is committed to establish an integrated and interoperable research infrastructure of language and cultural resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities.

To reach these goals each resource will be described using CMDI (Component Metadata Infrastructure). The metadata documents are indexed and made searchable using an advanced search platform, the CLARIN Search for language and Digital Heritage Resources (Zhang 2012). This search engine provides a solution for access to different types of language and cultural resources (information aggregation). It allows different groups of users to search across research data and locate relevant data for new insights. It features full-text search, faceted search and a first possibility for geospatial search.

Examples of datasets available (www.meertens.knaw.nl/cmdi/search):

- Soundbites, 1.983 records (www.meertens.knaw.nl/soundbites),
- Dutch Song Database, 243.129 records (www.liederenbank.nl).
- Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND), 268 records (www.meertens.knaw.nl/sand).

While the first steps have been made in combining big data and innovative search methods, systems like CLARIN still need to incorporate the tools and practices from existing humanities research. The question remains whether methodological innovation and advancing the modelling of humanities data and heuristics is better served by flexible small-scale research focused development practices. The paper will compare the data visualisation of small scale Virtual Research Environments (VRE), the Soundbites interface of the Meertens Institute, with the same dataset visualised in CLARIN. It will conclude with a reflection on these new methods and approaches in the humanities, with special attention to mining and interpreting big data more effectively.

References

Berry, David M. (ed.) (2012). *Understanding Digital Humanities* (Palgrave Macmillan, London).

David Bodenhamer (2012). One Place, Many Beliefs: Visualizing the Complexity of American Religion. In: *European Social Science History Conference 2012* (Glasgow).

Gregory, Ian (2012). GIS and Texts: Exploring Lake District Literature using GIS. In: *European Social Science History Conference 2012* (Glasgow).

Knorr-Cetina, K. (1999) *Epistemic Cultures: How the Sciences Make Knowledge* (Cambridge, Massachusetts: Harvard University Press).

Zeldenrust, Douwe (et al.) (2012). Exploring new ways of integrating heterogeneous spatial data and annotations. In: *European Social Science History Conference 2012* (Glasgow).

Zhang, Junte Marc Kemps-Snijders, and Hans Bennis (2012). The CMDI MI Search Engine: Access to Language Resources and Tools Using Heterogeneous Metadata Schemas, in:

Lecture Notes in Computer Science, 2012, Volume 7489, Theory and Practice of Digital Libraries, Pages 492-495.

Websites:

www.clarin.eu (Accessed April 4, 2013).

www.liederenbank.nl (Accessed April 4, 2013).

www.meertens.knaw.nl (Accessed April 4, 2013).

www.meertens.knaw.nl/cmdi/search (Accessed April 4, 2013).

www.meertens.knaw.nl/soundbites (Accessed April 4, 2013).

www.meertens.knaw.nl/sand (Accessed April 4, 2013).