Combining tailor made research solutions with big infrastructures: The speaking map of the Netherlands

Douwe Zeldenrust
Meertens Institute (Royal Netherlands Academy of Arts and Sciences)
The Netherlands
douwe.zeldenrust@meertens.knaw.nl

Marc van Oostendorp
Leiden University
The Netherlands
m.van.oostendorp@hum.leidenuniv.nl

Introduction

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. But there is growing scepticism concerning the value of these big infrastructures. In 2012, at the Digital Humanities conference, it was questioned if it is possible and desirable to have an infrastructure for the humanities (Bellamy, 2012). At the Cologne Dialogue on Digital Humanities 2012, this perception was even taken a step further. It was argued that digital infrastructures could be regarded as a dead end for digital humanities. According to this view methodological innovation and advancing the modelling of humanities data and heuristics is better served by flexible small-scale research focused development practices (Zundert, 2012).

While this discussion focuses on the question whether or not e-infrastructures are theoretically readily usable for specific research questions, the current challenge is much more concrete. It lies in catering for specific research needs and making the resources available for future and potential interdisciplinary research. This paper will focus on the possibility of creating tailor made solutions for researchers or Virtual Research Environments (VREs), while at the same time connecting, using and contributing to the big infrastructures. To make this tangible a use case will be presented: 'The speaking map of the Netherlands'. First of all this paper will give an overview of this project. Next it will go into detail concerning the connections with regard to the access, sharing and storage of the data. Finally, the paper will conclude with a reflection on future steps in connecting research data and tools to infrastructures.

Text for the speaking map of the Netherlands

The Meertens Institute, an institute of the Royal Netherlands Academy of Arts and Sciences, studies the diversity in language and culture in the Netherlands (Meertens). It possesses a large library and numerous (audio) collections. The institute has more than a thousand hours, or six weeks non-stop listening, of

audio recordings of dialects from all parts of the Netherlands. The recordings are of conversations between two or more people, without interference from the researcher. The institute started in 1950 to collect the data. In the eighties collecting stopped when the recordings were sufficiently spread out over the Netherlands. Since 2009 the dialects (in total 2216 recordings are available) can be found on the website of the Meertens Institute as the 'speaking map of the Netherlands' (Soundbites).

The Meertens Institute also has typescripts of 660 of the available recordings. These typescripts have been digitized and the collection contains in total more than 11,000 scans (Archives). Optical character recognition (OCR) has been performed on these typescripts and samples of the produced texts have been corrected. While this collection is digitized, it is not yet readily available for researchers. In July 2012 the Royal Netherlands Academy of Arts and Sciences funded the project 'text for the speaking map of the Netherlands' to provide open access to the typescripts and to facilitate research with the entire resource (including the audio files). The project started in September 2012 and it will run until May 2013.[1]

A tailor made solution

The project 'Text for the speaking map of the Netherlands' incorporates the lessons learnt from the construction of previous VREs (Berry et al., 2012). One of the key issues of constructing VREs is the implementation of tailor made interfaces for interaction between research questions, data, tools and infrastructure (Zeldenrust, 2011). To establish direct communication and to bridge the gap between research question and technical possibilities a small team has been formed. In conjunction with a phonologist, a programmer and the audio curator of the Meertens Institute, a dedicated interface for phonologic research has been designed.

The phonologic interface will be a web-based system. Its core is a MySQL database containing the metadata and the web locations of the audio files and the scans of the typescripts. The web application will present various ways of exploring the data. First of all, in the cartographic tradition of the Meertens Institute, the site offers the visual interface of the speaking map. It is a representation of the data using the geographic locations. Next, the web application provides access to the datasets using the metadata fields. This will allow scrolling through the data. And finally, using previous added keywords and the ORCs of the typescripts, a text search will also be available. These dedicated

interfaces need a relative small budget, are quick to set up and are able to serve as a stepping-stone for innovative research.

Connecting to a digital humanities infrastructure

The interface is specially designed for phonological research of the Dutch dialects. This field of study of research is currently flourishing. One could for instance perform a large-scale phonetic study into vowel quality and vowel length using the combination of sound and typescript, or research into word frequency in Dutch dialects. In addition, the collection offers a unique representation of the Netherlands that no longer exists. The conversations are about poor living conditions in rural areas, welfare during the Depression, the Second World War, local customs et cetera. This collection presents a rich resource for interested parties other than the traditional dialect researchers and is yet to be discovered by for example historians and ethnologist.

While the phonological interface makes the resource available for a specific field of research, the project intends to explore the full potential of the resource and to open it up for a wide variety of research possibilities. To reach this goal the resources will be made available through the Common Language and Resources Infrastructure (CLARIN). Each resource will be described using CMDI (Component Metadata Infrastructure) and will be assigned PIDs (Persistent Identifiers). To allow others to harvest our metadata records an OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) provider will serve the DCMI (Dublin Core Metadata Initiative) and the CMDI documents. The metadata documents are furthermore indexed and made searchable using an open source search platform.[2] It features full-text search, faceted search and geospatial search. CLARIN will handle the long term archiving of the data using the facilities of The Language Archive (TLA).[3] These big infrastructures are expansive and it takes time and mass to reach a critical usable level. In connecting the dataset to the CLARIN infrastructure it will be disseminated not only for phonology but also for other types of research.

Conclusion

In the introduction it was questioned if big infrastructures are potential platforms for methodological innovation. Some even take it a step further and state that big infrastructures could be regarded as a dead end and that flexible small-scale solutions serve humanities research better. In the case of project 'Text for the speaking map of the Netherlands' audio files, typescripts and metadata will be made available via a tailor made web interface. Big infrastructure CLARIN will

provide dissemination, storage and the possibility of combining the resources. The latter functionality is methodological not innovative, however, it may lead to new insights and knowledge. Using the standards of CLARIN also provides easy to use building blocks for future VREs. The conclusion is that at the moment both tailor made research solutions and big infrastructures are of value for research. This current opportunity is restricted to the use of resources; how we deal with tools in this respect is still a matter to be resolved.

References

Bellamy, Craig (2012). Opportunity and accountability in the 'eResearch push'. In: *Digital Humanities 2012, conference abstracts* (pp. 111–112).

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

Zeldenrust, Douwe and Marc Kemps-Snijders (2011). Establishing connections: Making resources available through the CLARIN infrastructure. In: *Supporting Digital Humanities 2011, Copenhagen, proceedings*. http://cst.ku.dk/sdh2011/papers (Accessed October 05, 2012).

Zundert, Joris (2012). If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities. In: *The Cologne Dialogue on Digital Humanities 2012, Controversies around the Digital Humanities, proceedings.* www.cceh.uni-koeln.de/events/CologneDialogue (Accessed October 05, 2012).

Archives
The Archives of the Meertens Institute, collection no. 62.

Websites
www.clarin.eu (Accessed October 05, 2012).

www.meertens.knaw.nl (Accessed October 05, 2012).

www.meertens.knaw.nl/soundbites (Accessed October 05, 2012).

www.mpi.nl/research/research-projects/the-language-archive (Accessed October 05, 2012).

Notes

[1] In December 2012 the Speaking Map received additional external funding. More than 240 hours of Dutch spoken in France and the USA will be added in 2013. Plans to add other collections are in the making.

[2] In this case SOLR is used.

[3] The data of the Speaking Map is used by the TLA in a pilot project concerning the archiving of data. This pilot started in November 2012.