Infrastructure for Other Humanities Disciplines

Infrastructure for Other Humanities Disciplines: Introduction

Jan Odijk and Arjan van Hessen

UiL-OTS, Utrecht University, j.odijk@uu.nl, a.j.vanhessen@uu.nl

25.1 Introduction

Though CLARIN originated in the linguistics and computational linguistics communities, CLARIN-LC (in particular CLARIN-NL) covers a lot of other Humanities disciplines. This is in part due to the bottom-up approach for subprojects for data curation and software demonstrators, and in part to the active policy to include these other disciplines, implemented with an interactive user survey and active 'evangelising' among researchers of all Humanities disciplines.

We will first provide a brief overall overview of the relevant data and software that resulted from CLARIN-LC (section 25.2), and then summarise the topics of the chapters of this part (section 25.3).

25.2 Work on Other Humanities Disciplines

Subdisciplines from the Humanities other than linguistics that were covered in CLARIN-LC include historical research, literary research, religion research, media research, social research, and philosophy:

Historical research The INTER-VIEWS application enables search in the IPNV interviews with war veterans, in part data curated in the INTER-VIEWS project. The WIP application enables search in parliamentary debates. WAHSP / BILAND[1] enables search in textual data of news media from the period 1863-1940 of the *Koninklijke Bibliotheek* and *Staatsbibliothek*

---

[1] These applications have in the meantime been superseded by Texcavator.

How to cite this book chapter:

Odijk, J and van Hessen, A. 2017. Infrastructure for Other Humanities Disciplines: Introduction. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 313–316. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.25. License: CC-BY 4.0

zu Berlin (see chapter 27). The VK application enables search in the collected works of Loe de Jong on the Netherlands in the Second World War (VK data). The ePistolarium application enables search in a corpus of 20,000 letters of scholars who lived in the 17th-century Dutch Republic (see chapter 26). The RemBench application enables searching and browsing for works of art, artists, primary sources and library sources related to Rembrandt (Rembench data, see chapter 28). DSS provides a tool chain and methodology for converting legacy datasets in the area of maritime history and a search application that enables search in maritime history, in particular in datasets related to recruitment and shipping in the East-India trade and in the shipping of the northern provinces of the Netherlands (DSS data). Nederlab, which is still under development, enables searching and analysing data from digitised texts spanning the full recorded history of the Netherlands, its language and culture. The Dutch Song Database (DSD) integrates four different datasets into a single database. It contains (meta-)data on 140,000 songs and their 15,000 sources (songbooks, pamphlets, field recordings, etc.) from the Middle Ages to the present day. The literary data also include EMIT-X: data and metadata from the Emblem Project Utrecht (EPU), which created a digital collection of 27 books of love emblems.

Literary research The Arthurian Fiction web application enables searching and browsing in data on mediaeval Arthurian narratives and the manuscripts in which they are transmitted throughout Europe (Arthurian Fiction data). COBWWWEB enables search in the Women-Writers Database and connected databases in women's literature, while NameScape enables searching for names and analysing their use in literary works (see chapter 30). BNM-I enables searching in a collection of textual, codicological and historical information about thousands of Middle Dutch manuscripts.

Religion research The PILNAR application enables search in a corpus of PILNAR pilgrims' narratives with Dutch texts written after 2000 (see chapter 31). The SHEBANQ application, already described in chapter 18, enables search in the SHEBANQ curated WIVU database containing the Bible text in Hebrew.

Media research Polimedia provides search in the minutes of the debates in the Dutch Parliament (Dutch Hansard) in combination with the databases of historical newspapers and ANP radio bulletins to allow cross-media analysis of coverage. AVResearcherXL enables the combined exploration of radio and television programme descriptions, television subtitles and general newspaper articles.

Social research MIGMAP enables searching and analysing data on migration flow between Dutch municipalities (see chapter 29).

Philosophy @PhilosTEI is a workflow for converting digital images into textual resources in TEI[2] format, and has specifically been applied to philosophical works (see chapter 32)

In addition, there are data from CLARIN data providers covering a wide number of disciplines. These data include the NISV Academia Collection and digital publications from Utrecht University Library, as well as digital publications from the National Library (KB).

25.3   Contents of Part IV: Infrastructure for Other Humanities Disciplines

The current book's part IV on infrastructure for other Humanities disciplines contains chapters for only a small sample of the resources described in section 25.2. We already referred to the chapters

---

2 TEI (Text Encoding Initiative) is a widely used standard for encoding textual resources supported by CLARIN.

that describe these resources. For the reader's convenience, we summarise the contents of part IV on infrastructure for other Humanities disciplines here and briefly describe the contents of each chapter:

Chapter 26 describes the *ePistolarium*, a virtual research environment for browsing and analysing a corpus of letters written by and sent to 17th-century scholars who lived in the Dutch Republic. It was developed in an independently financed project named *Circulation of Knowledge: A Web-based Humanities' Collaboratory on Correspondences and Learned Practices in the 17th century Dutch Republic (CKCC)*. The authors describe this project and provide an overview of the analysis methods that are available to the users of the ePistolarium, emphasising the role of Natural Language Processing techniques.

Chapter 27 claims that the human language technology that has been developed and used in the CLARIN demonstrator projects WAHSP and BILAND supports advanced forms of (multilingual) text mining in large datasets of newspapers. The authors argue that it is the massive processing of sources (pre-processed and offering a reliable critical text) – rather than the exhaustive analysis of a limited number of records – that will offer an added value to the historical sciences. The authors describe the development, use, and challenges of the WAHSP and BILAND text mining tools and their successor, Texcavator, to support distant reading in historical newspaper collections. They show how semantic text mining enables new and advanced forms of historical analysis based on case-studies focusing on the circulation of ideas and notions regarding drugs and eugenics during the first four decades of the 20th century.

Chapter 28 presents *RemBench*, a search engine for research into the life and works of Rembrandt van Rijn. RemBench combines the data from four different databases behind one interface using federated search technology. Metadata filtering is enabled through faceted search. RemBench enables art historians and other professionals interested in Rembrandt's period to find all information on Rembrandt that is available in online repositories in one application. The authors describe the user interface and results of its evaluation and claim that RemBench sets an example for search engines in the digital humanities.

Chapter 29 presents MIGMAP, software for the interactive mapping of socio-cultural phenomena in the Netherlands on the web. It demonstrates the possibilities that MIGMAP offers for the mapping of migration in the Netherlands across four generations. Both origin and dispersion of the population can be explored at the geographic levels of municipality, region, dialect area and province.

Chapter 30 presents NameScape, which enables researchers to carry out comparative literary onomastics on a large corpus of literary works. In comparative literary onomastics it is assumed that patterns and trends can be discovered in the way in which literary authors make use of proper names in their work. The NameScape project created a large corpus of literary works, made available tools to perform high-quality named entity recognition on literary material and tried to perform named entity resolution so as to determine whether names in literary works are plot internal or plot external. The data were made available in an environment in which the researcher can search and visualise search results.

Chapter 31 describes PILNAR, which created and opened up a corpus of Dutch pilgrim narratives for interested researchers. The growing number of narratives were collected and structured in a meaningful manner, providing a research tool that enables academics to work with this fascinating set of stories. The contribution takes a retrospective look at the construction of the PILNAR database and looks ahead to the possibilities of its results.

Chapter 32 addresses the problem of corpus building by developing an open source, web-based, user-friendly workflow from textual images to TEI, based on state-of-the-art open source OCR software and a powerful OCR post-correction tool developed earlier in

CLARIN-LC (TICCLops). The authors demonstrate the utility of the tool by applying it to a multilingual, multi-script corpus of important 18th- to 20th-century European philosophical texts, thus satisfying a basic pre-condition for the step towards e-research in philosophy.

Acknowledgements