CLARIAH in the Netherlands

Jan Odijk
Utrecht University
Utrecht, the Netherlands
E-mail: j.odijk@uu.nl

Abstract

I introduce CLARIAH in the Netherlands, which aims to contribute the Netherlands part of a Europe-wide humanities research infrastructure. I describe the digital turn in the humanities, the background and context of CLARIAH, both nationally and internationally, its relation to the CLARIN and DARIAH infrastructures, and the rationale for joining forces between CLARIN and DARIAH in the Netherlands. I also describe the first results of joining forces as achieved in the CLARIAH-SEED project, and the plans of the CLARIAH-CORE project, which is currently running.

Keywords: research infrastructure, Humanities, CLARIN, DARIAH, CLARIAH

1. Introduction

CLARIAH[1] is a series of projects[2] in the Netherlands that aim to design, construct and exploit a research infrastructure for humanities researchers as an integrated part of the European CLARIN[3] and DARIAH research infrastructures. In this contribution I will sketch the major characteristics of CLARIAH and the rationale behind them (section 4). In particular, I will describe the rationale behind joining forces for work on CLARIN and DARIAH in the Netherlands. I will start by sketching the changing landscape in the Humanities (section 2), and the context in which the project was conceived, both at the international and at the national level (section 3). I summarize the major conclusions in section 5.

2. The Humanities are Turning Digital

The amount of data that is available in digital form is increasing exponentially. This is true generally, but also for the Humanities in the Netherlands. It includes contemporary newspapers, journals, TV and radio broadcasts (texts of 1.5 million radio bulletins), new media (twitter, Facebook, etc.), but also historical newspapers (over 80 million articles, www.delpher.nl), books (over 170 k books from the 18th- 20th century) and magazines (over 1.5 million pages from the 18th and 19th century), etc.

The fact that the data are becoming available in digital form implies that they can be analyzed with digital techniques. In addition, the computer hardware enables this processing, basic analysis software is available, and advanced analysis techniques, inter alia natural language processing tools and applications, often yield sufficient quality to use them. Therefore, the so-called *Digital Turn* creates huge opportunities for the Humanities: it can broaden the empirical basis for its research, since digital techniques can analyze data in quantities that humans never could cover. It will therefore enable the investigation of existing research questions in new ways, create opportunities for investigating research questions that could not be addressed before, and for formulating and investigating completely new research questions.

However, this digital turn is not going to be easy! The reason is that this enterprise involves the entire spectrum of the analytics challenge of big data: it involves massively distributed data sources, both structured data and unstructured data, of varying complexity and quality (often with a lot of noise, partially incomplete, etc.). Large volumes of unstructured data come in multiple formats: audio, video, image, text, requiring formal (syntactic) and semantic interoperability. And the users of these digital data are globally distributed, and highly varied, across many Humanities disciplines, all speaking very different languages.

These problems are also familiar to modern software companies active in the area of text analytics, which attempt to analyze noisy digital texts and their metadata in a wide variety of formats and from a wide variety of social media platforms. These include language technology companies, but also their customers, and they include big players such as Microsoft and IBM. IBM explicitly recognizes the parallels: “The challenges faced by the Art & Humanities are highly representative and synergistic with the broader challenges IBM is solving across other industries – from law enforcement to health care and beyond”. In addition, the problem is familiar to

[1] The name CLARIAH is not unique to the Netherlands. Also Austria has a CLARIAH project with the goal to contribute to the CLARIN and DARIAH infrastructures. I will use CLARIAH for the project in the Netherlands, unless confusion might arise in which case I will use CLARIAH-NL for the Dutch and CLARIAH-AT for the Austrian project. CLARIAH is (in the Netherlands) an acronym for Common Lab Infrastructure for the Arts and Humanities.

[2] It consists of two projects which are named CLARIAH-SEED and CLARIAH-CORE. See below for more details.

[3] Note that several names and acronyms in this paper are hyperlinks to relevant web sites. One can access these hyperlinks by clicking on them in the digital version. Most PDF-renderings show the hyperlinks in a different font on paper.

public digital heritage organisations.[4]
Not only the data pose problems, there is also great variety among the intended users (humanities researchers): there are big differences among the humanities researchers in terms of their technical knowledge and expertise, and their willingness to embrace the digital techniques.
These problems make it necessary to set up a research infrastructure for the humanities to facilitate the Digital Turn, and CLARIAH aims to make significant contributions to this.

3. Context

CLARIAH aims to make contributions to research infrastructures for humanities researchers. An *infrastructure* is a set of usually large-scale basic physical and organizational resources, structures and services needed for the operation of a society or enterprise.[5] Typical examples are the railway network, the electricity network, or on a smaller scale, the availability of wireless internet through Eduroam at all Europe's educational premises. A *research infrastructure* is an infrastructure intended for carrying out research: facilities, resources and related services used by the scientific community to conduct cutting-edge research. Famous examples are the Chile Large Telescope and the CERN Large Hadron Collider. *Humanities researchers* include linguists, historians, literary scholars, philosophers, religion scholars, and others and include (in the CLARIAH context) researchers that usually are counted as social scientists, in particular political sciences researchers.
CLARIAH came into existence in an ecosystem of projects related to research infrastructures for the humanities. I will sketch the context at the international level (section 3.1) and at the national level (section 3.2).

3.1 International Level

At the international level two research infrastructures for humanities researchers are on the ESFRI Roadmap: CLARIN and DARIAH.
CLARIN (Common Language Resources and Technology Infrastructure) focuses on *language* and therefore provides facilities for digital language resources. Digital language resources include software and data. The data include textual data in natural language, databases about natural language (typological databases, lexical databases, dialect databases, etc.), and audio-visual data containing (written, spoken, signed) language. The latter include pictures of manuscripts, audio-visual data for language documentation and sign language description, interviews, radio and TV programmes. The software includes programmes for analyzing language in textual and audio-visual data, for enriching language data with a wide variety of linguistic annotations, and for searching in language data that contain these linguistic annotations. CLARIN considers language in all the functions it is used for, not only as an object of inquiry, but also as a carrier of cultural content, as a means of communication, and as a component of identity. CLARIN has set up an ERIC, a legal entity at the European level specifically set up for research infrastructures, which is hosted by the Netherlands. CLARIN ERIC currently has 17 member countries and one observer country.[6]
DARIAH (Digital Research Infrastructure for the Arts and the Humanities) aims to enhance and support digitally-enabled research and teaching across the Humanities and Arts. It is a network of people, expertise, information, knowledge content, methods, tools and technologies coming from various countries. DARIAH also set up an ERIC, which is hosted by France, and currently has 18 member countries.[7]
CLARIN and DARIAH are both distributed infrastructures. CLARIN is implemented in a network of CLARIN centres. These centres come in different flavours and include centres for general infrastructure services, centres for data and software services, and (virtual) knowledge centres.
DARIAH is organized in Virtual Competence Centres (VCCs) , of which there currently are four. Each VCC focuses on a particular theme: e-infrastructure, the liaison between research and education, the management of scholarly content, and advocacy, impact and outreach.
CLARIN and DARIAH are also both *virtual* infrastructures: most of their organizational units exist only digitally and are implemented in a distributed, international and often cross-disciplinary network of actual organisations, and both infrastructures provide their services mainly via the internet.
Some other international projects are also relevant in the context of CLARIAH.
Clio-Infra aims to research the long-term development of worldwide economic growth and inequality. It has set up a set of interconnected databases containing worldwide data on social, economic, and institutional indicators for the past five centuries. From the Netherlands, the IISH[8], Utrecht and Groningen Universities, and DANS are important players in this project.
EU-Screen(-XL) offers free online access to thousands of items of audio-visual heritage. It is a resource for educators, researchers and media professionals searching for new audio-visual content from across Europe. From

[4] Not surprisingly, CLARIAH received a lot of support from companies and organisations in the Netherlands dealing with similar problems. These include IBM, Microsoft, NOTaS, Furore, Knowledge Concepts, Telecats, STM, Arbor Media, Teezit, Tessella, and Brill Publishers
[5] Adapted from an earlier description on Wikipedia.
[6] See http://www.clarin.eu/content/national-consortia for an overview of the CLARIN ERIC member and observer countries.
[7] See http://dariah.eu/about/ourpartners.html for an overview of the DARIAH ERIC member countries.
[8] International Institute for Social History

the Netherlands the NISV[9] and Utrecht University play an important role in these projects

3.2 National Level

At the national level, a national project to contribute to the CLARIN infrastructure, called CLARIN-NL, was set up in 2009 and ran until 2015.

CLARIN-NL (Odijk 2014a,2014b,2014d) implemented the Netherlands part of the CLARIN research infrastructure in multiple certified CLARIN centres. It curated many digital data, and created many web applications specifically designed for the intended users, humanities researchers, inter alia with functionality for enriching data, searching in enriched data, analyzing the data and visualising the analysis results. The data and applications are easily accessible via the CLARIN-NL Portal, which contains faceted search options to search for data or applications that might be relevant to one's research, inter alia by facets for research discipline, tool task and language. CLARIN-NL also created 6 educational packages and 9 short movies and multiple screencasts to educate the targeted researchers on the opportunities that CLARIN offers them. CLARIN-NL focused on language but covered many humanities disciplines that use language resources, including linguistics, literary studies, history, political sciences, religion studies, philosophy, and media studies.

DARIAH in the Netherlands applied for funding for a national project, which was not awarded, so that the DARIAH activities in the Netherlands from 2009 through 2012 were very limited in scope.

Several other infrastructure projects provided a basis for CLARIAH.

Nederlab collects Dutch historical text corpora and enriches them with CLARIN-compatible metadata and all kinds of linguistic annotations. It provides access to these data via browse, search and analysis interfaces to a powerful search engine. It aims to support the longitudinal study of the Dutch language and culture.

HSN (Historical Sample of the Netherlands) offers a representative sample of life courses of 78,000 people born in the Netherlands (1812-1922) and a unique tool for research in Dutch history and demography.

CATCH (Continuous Access to Cultural Heritage) and CATCHplus aimed to make the (digital) collections of museums, archives and historical associations more accessible, and to Improve efficiency of heritage management.

4. Joining Forces

CLARIN and DARIAH in the Netherlands decided to join forces in 2011. This resulted in the CLARIAH-SEED project, which ran from 2012 through 2014. This project was intended for maintaining the consortium that had been set up covering essentially all the humanities faculties and institutes in the Netherlands, and for formulating a proposal for the 2014 revision of the Netherlands National Roadmap for large scale facilities.

CLARIAH-SEED resulted in a new CLARIAH proposal (which was awarded funding and led to the CLARIAH-CORE project, see section 4), it supported DARIAH activities by the Netherlands both at the national and the international level, it created several short movies, as well as the Dutch part of the course registry for Digital Humanities. It also resulted in a range of demonstrators:

- TrOVe (Transmedia Observatory), a search application to analyse the distribution of information throughout time across different media such as broadcast television, print media, social media and blogs. A mini research pilot was carried out using this application to investigate the relation between the media and Eastern European migration.
- OHT and OHT+ (Oral History Today), which supports the workflow of working with unstructured audio-visual content (esp. Oral History): archival search, browsing, playing fragments, making notes, visualization of patterns and publication.
- CLIO-DAP (CLIO Data Availability Policy), a demonstrator service for enhanced publications.
- Nederlab created a search and analysis application for historical Dutch text corpora. Nederlab was also used to carry out a mini research pilot project on author distribution in the literary journal *De Gids*.

CLARIAH-SEED also worked on data:

- **Stakingsdata** ('Strike Data'), in which labour conflicts are linked at the micro and macro level and visualized in dynamic historical maps.
- HLZ (HSN Links Zeeland), which combines HSN Zeeland and the LINKS database.
- Athena: a design was made for historical database on flora and fauna species in cultural and natural contexts for the Netherlands.

It was pretty natural for CLARIN-NL and DARIAH-NL to join forces. After all, there are many commonalities, which can be shared and need not be duplicated. Each has to ensure that data and tools can be discovered or found by researchers (visibility), can be accessed by them (accessibility), can be referred to in a persistent manner (referability, persistence), and are safely stored for long term preservation. Furthermore, it must be possible for researchers to apply tools and services to data in a seamless manner. Many of the techniques to achieve such functionality are common to CLARIN and DARIAH. CLARIN and DARIAH have are already been closely cooperating exactly for these reasons, inter alia in a European project such as DASISH. Combining force in the Netherlands will intensify such cooperation.

In addition, CLARIN and DARIAH are complementary in certain respects.. For example, CLARIN is strong on

[9] Netherlands Institute for Sound and Vision

aspects of the technical infrastructure, one essential component of a research infrastructure, while DARIAH is strong on the knowledge infrastructure, another essential component of a research infrastructure. By joining forces, CLARIN and DARIAH thus strengthen each other in these respects.

There are surely also many differences between CLARIN and DARIAH, but we believe that they are not significant enough to keep going our own ways, and any differences that must be overcome can only be overcome by close cooperation. For example, metadata are of crucial importance in both infrastructures. CLARIN uses CMDI (Broeder et al. 2010) as framework for metadata, while DARIAH allows multiple ways of organizing, creating, and representing metadata. In the CLARIAH context, efforts are on their way (and have partially already been implemented (Ďurčo & Windhouwer 2014)) to overcome these differences through Linked Open Data and mappings from and to various metadata frameworks via Linked Open Data in RDF format.

Of course, there are also more strategic reasons for joining forces. In the Netherlands, the Humanities have to compete for funding of research infrastructure projects with all other scientific disciplines (e.g. physicists, astronomers, biologists, etc.). With a single joint project proposal from the Humanities we have a better competitive position to be awarded such funds.

Though there are clear advantages of joining forces, there are also potential dangers . The biggest one is losing focus: CLARIN has a clear focus on language, but DARIAH is broader, and there are many rather different disciplines in the Humanities. In order to avoid this risk, CLARIAH focuses on three disciplines within the Humanities:

- Linguistics
- Social-economic History
- Media studies

And on the main data types used by these disciplines:

- (natural language) text
- Structured (often quantitative) data
- Audio-visual data

Since these data types are covered by the core disciplines, and since many data and tools for these core disciplines are also relevant for other Humanities disciplines, we expect that the limitation to these three core disciplines will not impede later extensions to other disciplines within the Humanities. In addition, we have ensured that desires and requirements from other Humanities disciplines are taken into consideration during the design and construction of the infrastructure.

The selection for these three disciplines is also motivated by the fact that they are forerunners in Digital Humanities in the Netherlands: linguistics builds upon CLARIN-NL and the Nederlab project (and many earlier projects such as the Spoken Dutch Corpus and the STEVIN programme (Spyns & Odijk 2013)). Social-Economic History builds on the Clio-Infra and HSN projects. Media studies builds on CLARIN-NL, CATCH and EU-Screen(-XL). And all build upon the results of CLARIAH-SEED.

5. CLARIAH-CORE

CLARIAH-CORE is the main project in the CLARIAH series. It has a budget of 12 million euro and runs from 2015 to 2018. It consists of work packages for each of the core disciplines, a work package for discipline-independent infrastructure services and data, and a work package for education and training, dissemination and outreach, IPR and ethical issues, and overall management. CLARIAH-CORE also aims to launch a call for research pilots.

The work package for general infrastructure services and data is working on

- Facilities for shared vocabularies
- Facilities for (mainly meta)data as Linked Data
- Search in the linked data (LD)
- Linking CMDI to LD, and vice-versa
- Access Control
- OCR / Text Correction and enrichment pipeline
- Standardization
- Performance and Availability

In the Linguistics work package, the goal is to support the linguist in each stage of a research project. These stages include:

- Data and Tool creation or collection, possibly via crowdsourcing
- Browsing / Searching for / selecting data and tools
- Enrich data with linguistic annotations
- Browse and search in (enriched) data + metadata
- Enrich, analyse, visualise search results
- Incorporate data / tools in CLARIAH
- Create enhanced publications

For each of these stages it has been inventoried what is needed, what is already available from earlier projects, and additions to and extensions of the existing functionality have been defined and are currently being implemented.

In the social-economic history work package databases at the *macro* (national/international), *meso* (trade unions, organisations) and *micro* (individual / family) levels are being linked. These databases have different histories, are structured in incompatible ways, and use different vocabularies. Integration of these databases is carried out using the Linked Data paradigm, and this integration will enable addressing research questions that require relating social-economic facts from different levels.

In the media studies work package the researchers will be supported by integrating improved versions of a range of independently developed applications in one Media Suite. The applications include

- CoMerDA, an aggregated search interface for audio-visual data
- AVResearcherXL for exploring audio-visual metadata in historical context
- TrOve (see above, section 4)
- DIVE: presentation of collection items in context and 'intuitive' browsing

- OHT (see above, section 4)

The research pilots are small research projects aimed at testing the infrastructure and/or specific parts of it. Such projects will lead to improved functionality, driven by concrete needs of humanities researchers working on one or a few closely related very concrete research questions. It may lead to successfully concluded research (reported on in publications), and/or to new requirements for the infrastructure or particular applications, services or data within the infrastructure. The work presented by (Odijk 2011, 2014c, 2015a) published in (Odijk 2015b, 2016) in the CLARIN-NL context could be considered a concrete example of such a research pilot avant la lettre. CLARIAH aims to launch the call for research pilots in 2016.

6. Conclusions

The Netherlands actively participates in CLARIN and DARIAH, and even has a leading role in CLARIN. The researchers in the Netherlands have been rather successful in securing funds for research infrastructures. The work on research infrastructures in CLARIN-NL and CLARIAH-SEED is yielding new research results, but there is still a lot that has to be done. We hope to strengthen research infrastructure for the humanities by joining the work for CLARIN and DARIAH in a single project, CLARIAH-CORE, which is now running for about 1.5 year. It is too early to evaluate the success of joining forces, but we hope that the experiences gained in the Netherlands (and in Austria) may benefit other countries in their decision to join forces or to let CLARIN and DARIAH go their own ways in their country.

7. Acknowledgements

This work was financed by the CLARIAH-CORE project, which is funded by the Netherlands Science Foundation (NWO).

8. References

[Broeder et al. 2010] Broeder, D., Kemps-Snijders, M., Uytvanck, D. Van, Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. (2010). A data category registry- and component-based metadata framework. In Calzolari, N., Maegaard, B., Mariani, J., Odijk, J., Choukri, K., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 43–47, Valetta, Malta. European Language Resources Association (ELRA).

[Ďurčo & Windhouwer 2014] M. Ďurčo and M. Windhouwer. From CLARIN Component Metadata to Linked Open Data. In *Proceedings of the third Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing* (LDL 2014). At LREC 2014, European Language Resources Association (ELRA), Reykjavik, Iceland, May 27, 2014.

[Odijk 2011] Odijk, J. (2011), "User Scenario Search", internal CLARIN-NL document, April 13, 2011. [docx]

[Odijk 2014a] Odijk, J. The CLARIN infrastructure in the Netherlands: 'What is it and how can you use it?', unpublished paper, Utrecht University [pdf] [URL]

[Odijk 2014b] Odijk, J. 'The CLARIN infrastructure in the Netherlands: Design and Construction', unpublished paper, Utrecht University [pdf] [URL]

[Odijk 2014c] Odijk, J. 'CLARIN: What's in it for Linguists?', Uilendag lecture, Utrecht, Mar 27, 2014. [pptx]

[Odijk 2014d] Odijk, J. 'CLARIN-NL: Major Results', in Proceedings LREC 2014, Reykjavik, May 2014, pp 2187-2193 [pdf] [URL]

[Odijk 2015a] Odijk, J. (2015), 'Using PaQu for language acquisition research', presentation at the CLARIN 2015 Conference, Wroclaw, Poland, 16 October 2015 [pptx] [pdf]

[Odijk 2015b] Odijk, J. (2015) 'Linguistic Research with PaQU'. *Computational Linguistics in The Netherlands Journal* 5, p. 3-14 [pdf]

[Odijk 2016] Odijk, J. (2016), 'A Use case for Linguistic Research on Dutch with CLARIN', accepted for K. De Smedt (ed.), 2016, *Selected Papers from the CLARIN 2015 Conference*.

[Spyns & Odijk 2013] Spyns, Peter and Odijk, Jan, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*, volume XVII of Theory and Applications of Natural Language Processing, chapter 13. Springer, Berlin, Germany. [URL]