Nederlab, towards a Virtual Research Environment for textual data

Mathijs Brouwer, Hennie Brugman, Matthijs Dröes, Marc Kemps-Snijders, Jan Pieter Kunst, Erik Tjong Kim Sang, Martin Reynaert, Rob Zeeman, Junte Zhang

Meertens Institute
Joan Muyskensweg 25
1096 CJ Amsterdam
The Netherlands

E-mail: matthijs.brouwer@meertens.knaw.nl, hennie.brugman@meertens.knaw.nl , matthijs.droes@meertens.kanw.nl , marc.kemps.snijders@meertens.knaw.nl , janpieter.kunst@meertens.knaw.nl, reynaert@uvt.nl , erik.tjong.kim.sang@meertens.knaw.nl , rob.zeeman@meertens.knaw.nl, junte.zhang@meertens.knaw.nl

Abstract

The Nederlab project aims to bring together all digitized texts relevant to Dutch national heritage, the history of Dutch language and culture (c 800 – present) in one user friendly and tool enriched open access web interface. At the time of writing the Nederlab environment holds over 37 million documents and provides various ways of accessing, filtering and visualizing result sets. It derives many ideas, services and technologies from subsequent layers of the Collaborative Data Infrastructure as presented by the High Level Expert group on Scientific Data.

Keywords: Nederlab, Virtual Research Environment, Collaborative Data Infrastructure, Dutch diachronic research corpus, Meertens,

1. Introduction

In 2010 the Collaborative Data Infrastructure was presented [*High Level Expert Group on Scientific Data, Riding the wave, 2010.*] as a future framework for scientific infrastructure development. In this framework essentially three layers (data generators/users, community support services and common data services) are distinguished showing close interaction between one another. Many of the currently existing infrastructure initiatives or community oriented projects in the humanities domain can easily be placed within this framework and generally exhibit a strong focus on either the Community Data Services or Common Data Services layer. Projects such as CLARIN or META-SHARE primarily focus on providing Community Support Services such as metadata repositories services or reusable processing services, e.g. NLP services. The currently running EUDAT project can be best positioned at the level of Common Data Services providing cross-domain solutions such as safe replication, simple storage and data staging services. While all of these initiatives provide useful services the end user focus remains limited. Although many of the projects do contain user-oriented components these are generally targeted at infrastructure internal participants or, at best, focus on providing accessibility options for the specific user community involved in the research questions/use cases underlying the project proposals.

As yet, however, no comprehensive research corpus has been established bringing together corpora from various sources accompanied by relevant tools for processing, searching and analyzing the data, mainly for practical reasons. The available corpora are housed in different places, with different metadata, and cannot be searched simultaneously. Advanced software tools necessary for the analysis of the texts are available, but are in most cases specific to a particular goal rather than generally applicable. The potential of the available technological means and scientific standards of verifiability and replicability is far from being exploited. The vision of an integrated Virtual Research Environment is only slowly starting to become reality.

2. Nederlab

The Nederlab project aims to bring together all digitized texts relevant to Dutch national heritage, the history of Dutch language and culture (c. 800 -present) in one user-friendly and tool-enriched open access web interface, allowing scholars to simultaneously search and analyze data from texts spanning the full recorded history of the Netherlands, its language and culture.

The project builds on various initiatives: for corpora Nederlab collaborates with the scientific libraries and institutions, for infrastructure with CLARIN (and potentially CLARIAH), for tools with eHumanities programmes such as Catch and IMPACT. Nederlab has the added value of creating a user-friendly infrastructure for researchers, which will promote cooperation and synergy as well as the formulation of new, often interdisciplinary, research questions.

3. Data generators

Data within Nederlab is currently obtained from the National Library of the Netherlands(KB) and the Digital Library of Dutch Literature(DBNL). In quantitative terms these are expected to provide the bulk of Nederlab metadata and data records, but in time data from other data providers will be incorporated as well. Within the CLARIN project considerable experience has been gained with harvesting and harmonizing metadata descriptions from various sources using the CMDI

metadata framework [Broeder 2011][Broeder 2010][CMDI]. Within Nederlab the CMDI approach has been selected as the preferred method of metadata delivery as this provides the possibility of automated mapping and ingest procedures. However, in practice the number of data providers following the CMDI approach is still limited and although many data providers, such as the KB and DBNL, have expressed interest in becoming CMDI data providers, metadata and data are currently delivered in various formats. Metadata formats encountered from these data providers include DIDL[1], ALTO[2] and idiosyncratic TEI formats. To handle these formats customized import processes are necessary to ingest metadata and data into Nederlab.

Another group of data providers in Nederlab consists of the editorial team carrying out quality assessment checks and, where necessary, modifications to the metadata documents. A separate editorial environment is being built into Nederlab providing quality assessment, metadata modification and data harmonization tools to support evaluation and validation procedures. After successful validation and approval, the submitted research data is synchronized with the Nederlab corpus.

Finally, it is envisaged that may upload their own data sets into Nederlab. These data will initially be available in their private workspaces and may be shared with others, but may also be submitted to the editorial board for acceptance within the public Nederlab environment.

4. Community Support Services

At the level of Community Support Services incoming metadata and texts are indexed to be able to provide serendipitous and focused search facilities for the user interaction side of Nederlab[Zhang 2012]. Given the large amounts of data available in Nederlab multiple indices are necessary here as Nederlab relies on an aggregated search strategy rather than a distributed one as pursued by the CLARIN project. An aggregated search strategy delivers a higher level of control of the metadata and data within the Nederlab context and, at the data level, does not rely on the individual search engine capabilities provided by the data providers. All metadata documents are time stamped and placed under linear version control to track the change history of these documents.

Since part of the Nederlab corpus consists of rather low quality newspaper data that have been automatically digitized through Optical Character Recognition (OCR) techniques it was deemed necessary to raise the quality of these digitized texts. For this, a customized version of TICCL (Text-Induced Corpus Cleanup) [Reynaert 2010] was used to reduce the amount of spelling variation introduced by the OCR process.

Furthermore, a portion of the data is automatically enriched with POS tags and Named Entities information for more linguistically oriented search strategies. Here, FROG [van den Bosch 2007] is used for tagging/named entity recognition and Blacklab[3] is used as a corpus retrieval engine.

A Lexicon service provided by the Institute of Lexicology[4] is added to the mix to be able to expand elementary search terms to historical variants.

Most of the services described above have previously been made available in the context of the CLARIN-NL project thus laying the foundation for reuse in this project.

Besides these more generic services, the Nederlab project also requires project specific services, for example to store workspace data. Workspaces are used to allow researchers to reproduce search results at later stages workspaces by storing user queries. Also a Nederlab specific Broker service[Brouwer 2013] was introduced to allow separation of front- and backend development and to orchestrate the various Nederlab service requests.

5. Common Data Services

Common Data Services typically have strong cross-domain features such as (persistent) storage, identification or execution environments. At this level the Nederlab project builds upon compute and data facilities provided by SURFSara[5]. In particular the HPC cloud environment[6] provides a flexible solution for rapidly deploying multiple virtual machine instances when the need arises. This has proven to be useful in the try out phases of the project when ingest and processing procedures are still in flux and in production phases where large amounts of data are to be processed rapidly. Since the HPC environment also provides storage facilities Nederlab data may reside within this environment thus alleviating the need for external storage during the construction phase.

6. Users

The Nederlab environment aims to provide end users access to large amounts of diachronic data relevant to Dutch language and culture. At the time of writing, Nederlab already contains over 37 million documens and the number is expected to rise further in the near future as new data sets are incorporated into the environment. To facilitate the need of high quality data the ingest process is supervised and monitored by an editorial team. An integrated editorial environment supports the

[1] http://xml.coverpages.org/mpeg21-didl.html

[2] http://www.loc.gov/standards/alto/techcenter/structure.php

[3] https://github.com/INL/BlackLab

[4] http://www.inl.nl/

[5] https://www.surfsara.nl/

[6] https://www.cloud.sara.nl/

editorial team in performing quality assessment checks, modifications and data alignment tasks.

For non-registered users the Nederlab environment provides powerful search tools focused on supporting both serendipitous and focused search. The user interface is specifically designed to support user in an usable and intuitive manner. Especially since the number of search results found can be extremely large, the user is no longer only served with standard result set lists, but needs to gain an understanding of the scope of the result set as a whole. On obvious view on such result lists is a diachronic one, but the richness of metadata opens up possibilities of other visualizations as well, such as genre, geographical or even gender or age distributions when authors are concerned. It also becomes possible to visualize combinations of these dimensions, such as for example the distribution of genres over time or the development of male and female authors over time. While these visualizations do not only provide visual means for displaying they also provide new ways of filtering and searching as the visualizations can be made navigable.

Registered users have the added benefit of being able to store their queries in their own workspace and manipulate them further at any give time in the future by adding or removing resources or further filtering of the results set. Also, the results remain reproducible, as each query is stored with associated time stamp information. This has the added benefit that modifications introduced by the editorial team at later stages can be suggested to the end users as extensions/updates to their search process. Finally, the user may compare multiple result sets to detect differences, patterns or anomalies.

7. Conclusion

The Nederlab environment builds upon the subsequent layers of the Collaborative Data infrastructure and is creating a Virtual Research Environment drawing upon resources from various data providers and considering the full resource life cycle. It has a strong end user focus aimed at supporting the end user in their data selection tasks and offering multiple data perspectives. It is envisaged that from this, the user can be further supported in more advanced data analysis tasks that will be specified in close collaboration with several independent research projects. And although it is realized that not *all* analysis tasks can be supported the Nederlab environment attempts to build a basic framework with relevant data for the Dutch language and culture domain, advanced search features and the opportunity to create different views on the selected data sets. For these data sets export options will be made available to support further analysis for cases Nederlab does not provide for.

8. Acknowledgements

The Nederlab project described in this paper has been made possible through projects grants from the Netherlands Organisation for Scientific Research (NWO), CLARIN-NL and CLARIAH.

9. References

Broeder, D., Schonefeld, O., Trippel, T., Van Uytvanck, D., & Witt, A. (2011).A pragmatic approach to XML interoperability — the Component Metadata Infrastructure (CMDI). *Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies, 7.*

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., & Zinn, C. (2010). A data category registry- and component-based metadata framework. *In N. Calzolari, B. Maegaard, J. Mariani, J. Odjik, K. Choukri, S. Piperidis, M. Rosner, & D. Tapias (Eds.), Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) (pp. 43-47). European Language Resources Association (ELRA*).

Brouwer, M. et all. Providing searchabilty using broker architecture on an evolving infrastructure *Submitted for LREC 2014*

Bruin de, M. Kemps-Snijders M., Kunst J.P., Peet van der, M., Zeeman R., Zhang J. (20120) Applying CMDI in real life: the Meertens case. *Workshop 'Describing Language Resources with Metadata', LREC 2012, Instanbul, 22-05-2012*

CMDI, ISO TC 37 SC 4 work item for ISO 24622)

Dima, Emanuel, Christina Hoppermann, Thorsten Trippel and Claus Zinn (forthcoming): "A Metadata Editor to Support the Description of Linguistic Resources". *Accepted to LREC 2012, the 8th International Conference on Language Resources and Evaluation*.

M. Gavrilidou, P. Labropoulou, S. Piperisid, M. Monachini, F. Frontini, G. Francopoulo, V. Arranz, V. Mapelli. A Metadata Schema for the Description of Language Resources (LRs), *International Joint Conference on Natural Language Processing, Chiang Mai / Thailand*

Marc Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg, Sue Ellen Wright (2009) ISOcat: remodelling metadata for language resources. nternational *Journal of Metadata, Semantics and Ontologies*

Reynaert M. (2008)Non-Interactive OCR Post-Correction for Giga-Scale Digitization Projects. *A. Gelbukh (Ed.), Proceedings of the computational linguistics and intelligent text processing 9th international conference (pp. 617-630). Berlin/Heidelberg: Springer*

Reynaert, M.W.C. (2010). Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition vol. 14 (2010) nr. 2 p.173-18.*

Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, Leuven, Belgium, pp. 99-114.*

Zeldenrust, D.A. & M. Kemps-Snijders. (2011) Establishing connections: Making resources available through the CLARIN infrastructure. *Supporting Digital Humanities 2011, Answering the Unaskable. Kopenhagen : [s.n.], 2011*

Zhang J. (2012) Supporting Serendipitous and Focused Search. *Proceedings of the 2nd European Workshop on Human-Computer Interaction and Information Retrieval, Nijmegen, The Netherlands, August 25, 2012, CEUR Workshop Proceedings, ISSN 1613-0073. [S.l.] : [s.n.], 2012, pp. 79-82.*

Zhang, J. and Kemps-Snijders, M. and Bennis, H.J. (2012) The CMDI MI Search Engine: Access to Language Resources and Tools Using Heterogeneous Metadata Schemas. *In: Theory and Practice of Digital Libraries - Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings, (pp. 492-495).*