PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/162489

Please be advised that this information was generated on 2017-11-12 and may be subject to change.
Nederlab: Towards a single portal and research environment for
diachronic Dutch text corpora

Hennie Brugman1, Martin Reynaert2, Nicoline van der Sijs2, René van Stipriaan1,
Erik Tjong Kim Sang1, Antal van den Bosch2, Jan Pieter Kunst1, Rob Zeeman1,
Dieuwertje Kooij1, Ineke Brussee1, Matthijs Brouwer1, Marc Kemps-Snijders1, Hans Bennis1
Meertens Institute1, CLST / Radboud University2, TiCC / Tilburg University3
The Netherlands

1. Introduction

Many digital Dutch text collections, from the eighth century to the present day, are available for digital humanities researchers. These texts reflect the dynamic development of Dutch language and culture. However, these collections typically are hosted by different institutions, are described with different metadata, and cannot be searched simultaneously. [FP]

The Nederlab project is building a comprehensive research corpus that brings together collections from various sources in the Netherlands and Flanders. It also brings together relevant tools for processing, searching and analyzing these data, in one virtual research environment that primarily supports historians, literary scholars, and linguists. The specific focus of Nederlab is on patterns of change over time and space.

Collections that are aggregated and integrated into Nederlab are processed using a carefully designed Nederlab collection pipeline (see Section 3). [FP] To all texts we apply sentence splitting, tokenization, lemmatization, part of speech tagging, and named entity recognition.

This uniform annotation treatment enables us to obtain the necessary statistics for word tokens and types, lemmata and PoS tags across all incorporated corpora. Federated Search (Stehouwer et al., 2012) across non-uniformly annotated corpora simply cannot provide this.

Metadata, text and linguistic annotations are indexed in one large, powerful search index. [FP] Section 2 of this abstract introduces the Nederlab project. Section 3 discusses how we integrate collection data into Nederlab, while Section 4 focuses on how we subsequently search and exploit these data. Section 5 presents a number of additional web services that we built, and that are valuable services in other contexts as well.

2. The Nederlab project

The Nederlab project started in 2013. It aims to bring together all digitized texts relevant to Dutch national heritage, in particular to the history of Dutch language and culture, 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. The project ties in with other major projects and initiatives: for collections Nederlab collaborates with academic libraries and institutions in the Netherlands and Flanders, for infrastructure with CLARIN (Odijk, 2010) and CLARIAH, for tools with eHumanities programmes such as NWO CATCH and IMPACT. The Nederlab project is currently about halfway. We have a solid collection pipeline in place for processing metadata, text content, and vocabularies, and have tested and applied this pipeline on three collections. The Nederlab index now contains 13.5 million searchable documents. We developed software for end users, the Research Portal, and have produced scripts and interactive tools for our back-office processes. The main focus for the rest of project period is on search and analysis of linguistic annotations on a massive scale, drawing on results of the current CLARIAH project MTAS, and on providing analytic tools in the context of scientific use cases. Efficiently adding new collections to Nederlab is also a major point of attention. In this Nederlab acts as a ‘user’ to the CLARIAH project PICCL in which a corpus building workflow is further being developed (Reynaert et al., 2015).

3. The Nederlab collection pipeline

Within the CLARIN project considerable experience has been gained with harvesting and harmonizing metadata descriptions from various sources using the CMDI metadata framework (Broeder et al., 2010). Within Nederlab the CMDI approach in principle has been selected as the preferred method of metadata delivery as this provides better chances for automated mapping and ingest procedures. However, in practice the number of collection providers following the CMDI approach is still limited. [FP] In almost all cases Nederlab has to deal with customized import processes to ingest metadata and data into the portal. These are the steps in our per-collection workflow:

Acquisition and IPR arrangements Early in the discussions with potential collection providers Intellectual Property Rights are addressed. We aim for a simple, standard contract and explicit agreement on what access policies we will implement to enforce this contract.

Quality Assessment and collection description We systematically collect information about each collection. This

1We mark by [FP] the parts of our abstract that are to be expanded in the Full Paper.
2Online at http://www.nederlab.nl/
3http://www.clariah.nl/en/
information is used to support internal data processing and
curation, and to inform end users about status and quality
of the collection’s data.

**Metadata mapping** We designed a fixed metadata
schema for the four basic Nederlab resource types (Titles
- in the sense of a work; Dependent Titles - only existing
as part of another Title; Series - like newspapers or period-
icals; and Persons - most importantly Authors) and repres-
ented and documented this schema using the CMDI frame-
work. Metadata of incoming collections is either mapped
to Nederlab metadata, or imported as ‘collection specific’
metadata, or ignored. This mapping is executed by the Neder-
lab editorial staff with a tailor-made metadata mapping
tool.

**Metadata conversion** For each collection custom con-
version scripts are written. The converted metadata is
stored in a project-internal relational database. This
database is then used for all curation tools and for the in-
dexing process.

**Text extraction and conversion** Text content is extracted
from the collection resources, sometimes with a different
granularity than the original (e.g. each newspaper article
is extracted as a separate Nederlab title). It is then con-
verted to the FoLiA XML format (van Gompel and Rey-
naert, 2013) and stored on a project internal FoLiA store.
Subsequent text enrichment or indexing processes are per-
formed on the FoLiA documents in the store.

**Curation by editorial staff** To facilitate the need for high
quality data the ingest process is supervised and monitored
by an editorial team. Metadata is manually curated. Au-
thors and titles are ‘thesaurised’: in a semi-automated pro-
cess authors and titles from newly integrated collections are
linked to already existing authors and titles. [FP]

**Text processing and enrichment** Since part of the Neder-
lab corpus consists of rather low quality 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 cus-
tomized version of TICCL (Text-Induced Corpus Cleanup)
(Reynaert, 2010) is used to reduce the amount of spelling
variation introduced by the OCR process. Furthermore,
the data is automatically enriched with lemmata and POS
tags and Named Entities labels by means of Frog (Van
den Bosch et al., 2007). Frog is developed for modern
Dutch, and the results for historical variants of OCR-post-
corrected Dutch vary from reasonable to mediocre; we are
working on improving this. [FP]

**Indexing** Incoming metadata and texts are periodically
indexed. This index allows the user to efficiently search
text and metadata, and to select a personal research cor-
pus out of the main corpus. Currently, we are working on
the next generation of our indexing and search soft-
ware, that, in addition, is capable of searching for com-
plex patterns of multi-layered linguistic annotations. We
closely collaborate with the Institute for Dutch Lexicology
(INL), who provide the corpus back-end BlackLab[4] and in-
tend to use front-end WhiteLab[5], further being developed
in the CLARIN-NL project OpenSoNaR-CGN, the sequel
to (Reynaert et al., 2014).

4. Virtual Research Environment

Since March 2015 a beta version of the Nederlab Research
Portal is online. It provides access to the first three of many
collections. One collection, the DBNL (Digital Library of
Dutch Literature) collection, contains high-quality tran-
scribed texts and extensive, well-curated metadata. The
second collection, Early Dutch Books Online, contains
historic digital texts digitized by means of OCR. For this
collection, Nederlab contains two alternative text versions
per paragraph: the original OCR text and an automatically
OCR post-corrected version for which (Reynaert, 2014) of-
fers a description and an evaluation. The third collection
was chosen partly to test scalability issues: the KB’s news-
paper collection up to 1900. All researchers have access
to the Nederlab search interface. They can select the way
in which their search results are represented: as a pageable list
of result snippets, as keyword-in-context concordance or,
visually, as a time distribution graphic showing the num-
bbers of matching documents over time. [FP] To enjoy all
the functionalities, users have to log in with a user account
in the CLARIN federation. Authorized users have three
additional benefits over non-authorized users: they are al-
lowed to inspect more text content, they are able to store
their queries as virtual research collections in their personal
workspace, and they have access to a growing number of
analytical tools to work on these virtual research collec-
tions. Currently, an initial set of analytical tools is available.
These tools are mainly focused on exploring and visualiz-
ing metadata such as distributions over genre, locations, or
gender and age information of authors. It also is possible
to visualize combinations of these dimensions. These visu-
alizations do not only provide visual means for represent-
ing metadata, they also provide new ways of filtering and
searching as the visualizations can be made navigable. At
this time we only provide for document count, not for term
count. However, we are working on a new search index that
allows for term counts as well. The next step is to expand
the set of available analytical tools. As a first step, we will
make the functionality of WhiteLab available. [FP]

5. Additional web services

Nederlab makes use of a number of web services that in
principle can also be used on their own, in other contexts.

**Lexicon service** The INL contributions to Nederlab in-
clude a historical Dutch lexicon. This lexicon is accessible
using a RESTful web service.9.

---

4https://github.com/INL/BlackLab/wiki
5https://github.com/TiCCSoftware/WhiteLab
6http://www.dbnl.org/
7http://www.earlydutchbooksonline.nl/nl/edbo
8http://www.delpher.nl/nl/kranten/
9http://sk.taalbanknederlands.inl.nl/LexiconService/
User annotations and Alexandria  Huygens ING builds Alexandria, a repository for text and annotations for Nederlab. It will be used to store and retrieve user generated annotations for all kinds of objects and object segments in Nederlab.

R based visualization service  We chose to base Nederlab visualizations on an separate web service that is based on the R open source software environment for statistical computing and graphics. This allows us, and potentially end users as well, to plug in custom R modules in the future.

6. Conclusion

Half way through the Nederlab project we have versions of most required components in place. We cover the whole trajectory from selecting and evaluating source collections all the way to generating statistical analyses over these collections in the context of all the other collections. Some of these components are still rudimentary, most of them need further development. We have tested all of this by processing three very different collections. We make these collections available in our Research Portal in a homogeneous and useful way, although we have not reached the full potential. We have gained insight in the processes and technology we need and in scalability issues. At this stage, Nederlab is constructed as an extensible framework. It can be extended by adding a variety of scholarly tools, as well as more collections, both during and after the remaining project period.

7. Acknowledgements

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

8. References


https://www.r-project.org