Common Lab Research Infrastructure for the Arts and Humanities

CLARIAH

A Research Infrastructure for the Humanities in the Netherlands

Jan Odijk
Vienna, 2015-04-09
The Humanities are turning Digital
European Context
National context
CLARIAH CORE
Conclusions
The Humanities are turning Digital

- European Context
- National context
- CLARIAH CORE
- Conclusions
THE DIGITAL TURN

• Amount of digital data is increasing exponentially
Also in the humanities:

- Contemporary newspapers, journals, tv, radio,
- New media: blogs, twitter, facebook, youtube,
- Historical Newspapers: > 80 M articles
  - www.delpher.nl
- Books: > 170 k books 18\textsuperscript{th} – 20\textsuperscript{th} century
- Magazines: > 1.5 m pages 18\textsuperscript{th} – 19\textsuperscript{th} century
- Texts of 1.5 m radio bulletins
• Digital data
  – ➔ Digital techniques can be used to analyze them
  – ➔ Hardware is enabling it
  – ➔ Basic Software is available
  – ➔ Advanced analysis techniques already often yield workable quality
THE DIGITAL TURN

• Huge opportunities for the humanities
  – Broaden empirical base of the research
  – Investigate existing research questions in a new way
  – Investigate research questions that could not be addressed before
  – Formulate new research questions
• But it is not going to be easy!
• Involves the entire spectrum of the analytics challenge of big data:
  – Massively, distributed data sources – structured, unstructured, complex, noisy and incomplete
  – Large volumes of unstructured data in multiple formats – audio, video, text, image – requiring semantic interoperability
  – Globally distributed and varied users – across disciplines – speaking very different languages
But with broad interest from industry

- Text analytics companies
  - Big customers of such companies, e.g. Philips
- Language technology companies
- Publishers
- Big players such as Microsoft and IBM
- IBM: cf. Watson
  - 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

- And public Digital Heritage organisations
INDUSTRIAL INTEREST
• But it is not going to be easy
  • Big differences within humanities
    – Technical knowledge and expertise
    – Willingness to embrace digital techniques
    – the digital turn must be facilitated by research infrastructures for the humanities
The Humanities are turning Digital

European Context

National context
CLARIAH CORE
Conclusions
• Infrastructure:
  – (Usually large-scale) basic physical and organizational resources, structures and services needed for the operation of a society or enterprise
  – Railway network, road network, electricity network, ...
  – eduroam
• Research infrastructure
  • Infrastructure intended for carrying out research: facilities, resources and related services used by the scientific community to conduct top-level research
  • Famous ones: Chile large telescope, CERN Large Hadron Collider
• humanities researcher
  • Linguists, historians, literary scholars, philosophers, religion scholars, ....
  • And a little bit in the social sciences: e.g. political sciences researchers
EUROPEAN CONTEXT: CLARIN

- **CLARIN** (Common Language Resources and Technology Infrastructure)
- Digital language-related resources
  - Data in natural language (texts, lexicons, grammars)
  - Databases about natural language (typological databases, dialect databases, lexical databases, ...)
  - Audio-visual data containing (written, spoken, signed) language (e.g. pictures of manuscripts, av-data for language description, description of sign language, interviews, radio and tv programmes, ...)
• Language in various functions
  • As object of inquiry
  • As carrier of cultural content
  • As means of communication
  • As component of identity
• The CLARIN infrastructure
  • Is distributed: implemented in a network of CLARIN centres
  • Is virtual: it provides services electronically (via the internet)
• Prepared by CLARIN preparatory project (2008-2011)
  – coordinated by Utrecht University
  – From Feb 2012 coordinated by CLARIN-ERIC, hosted by the Netherlands
  – ERIC: a legal entity at the European level specifically for research infrastructures
  – Other ERIC members: Austria, Bulgaria Czech Republic, Denmark, Dutch Language Union, Estonia, Germany, Lithuania, Norway, Poland, Portugal, Sweden and growing
• DARIAH (Digital Research Infrastructure for the Arts and Humanities)
  – aims to enhance and support digitally-enabled research and teaching across the humanities and arts
  – is a network of people, expertise, information, knowledge, content, methods, tools and technologies coming from various countries
EUROPEAN CONTEXT: DARIAH

• Prepared by DARIAH preparatory project (2008-2011)
  – coordinated by DANS (NL)
  – From Aug 2014 coordinated by DARIAH-ERIC, hosted by France
  – Founding DARIAH members:
    • Austria, Belgium, Croatia, Cyprus, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, The Netherlands, Serbia and Slovenia
  – And growing: Switzerland
• The Humanities are turning Digital
• European Context

➤ National context
• CLARIAH CORE
• Conclusions
• CLARIN-NL: 2009-2015

• Main results
  – RI implemented in multiple certified CLARIN centres
  – Many curated digital data
  – Many user-friendly web applications
    • Multiple functions: search, enrich, analyze, visualize, ...
Main results (cont.)

- Easily accessible via CLARIN-NL Portal
  - Faceted search for services and data
  - Covering many humanities disciplines
    - Linguistics, literary studies, history, political sciences, religion studies, philosophy, communication and media studies, ...
- With 6 educational packages and other educational material incl. 9 short movies and multiple screencasts
• **Faceted search**
  – By research discipline
  – By tool task
  – And other facets
  – Select result, introduction + link to application

• **Examples**
  – *Syntax, parsing* => *PaQu*
  – *Art History* => *RemBench*
NATIONAL CONTEXT: DARIAH-NL

• DARIAH-NL not awarded funding in 2008
• Joint CLARIN-NL / DARIAH-NL proposal in 2011: CLARIAH
  – Common Lab Infrastructure for the Arts and Humanities
• CLARIAH-SEED project
  – 2012-2014
• Why together?
• Plus
  – Many commonalities
  – Partially complementary
    • CLARIN strong on technical infrastructure
    • DARIAH strong on knowledge infrastructure
• Minus
  – But there are also many differences
  – Danger of losing focus
Focus on 3 disciplines
- Linguistics
- Social economic history
- Media studies

And corresponding main data types
- (natural language) text
- Structured (mainly quantitative) data
- Audio-visual data
• Linguistics
  – Building on CLARIN-NL
  – And Nederlab
• Social economic history
  – Building on Clio-Infra and HSN
• Media studies
  – Building on CLARIN-NL, CATCH, and EU-Screen(-XL)
NATIONAL CONTEXT: CLARIAH-SEED

• **Nederlab**
  – Independent project, co-financed by CLARIN-NL and CLARIAH-SEED
  – Dutch historical text corpora with CMDI metadata and all kinds of linguistic annotations
  – Access via browse, search and analysis interface
  – **Initial version** (limited access)
  – Aimed to support the longitudinal study of the Dutch language and culture
NATIONAL CONTEXT: CLARIAH-SEED

- **Clio-Infra**

- aims to research the long-term development of worldwide economic growth and inequality

- Has set up a set of interconnected databases containing worldwide data on social, economic, and institutional indicators for the past five centuries
NATIONAL CONTEXT: CLARIAH-SEED

- Clio-Infra

- Examples
  - GDP per capita, 1270-2000
  - Average years of education per capita, 1880 and 2000
  - Income per capita and life expectancy in China, Congo and NL, 1800-2007
• **HSN**
  – Historical Sample of the Netherlands

• offers a representative sample of life courses of 78,000 people born in the Netherlands (1812-1922)

• a unique tool for research in Dutch history and demography.
• **EU-Screen(-XL)**
  - offers free online access to thousands of items of audiovisual heritage
  - a resource for educators, researchers and media professionals searching for new audiovisual content from across Europe
• **CATCH** (Continuous Access to Cultural Heritage) and **CATCHplus**
  – make the (digital) collections of museums, archives and historical associations more accessible.
  – Improve efficiency of heritage management
NATIONAL CONTEXT: CLARIAH-SEED

• CLARIAH-SEED: Major Results
  – New CLARIAH proposal, awarded funding
  – NL DARIAH activities in EU and NL
  – Dissemination
    • Website
    • Short movies
    • Dutch Overview of the Digital Humanities
      – Course and project registries
• CLARIAH-SEED: Major Results
  – Demonstrators
    • TrOve (Transmedia Observatory)
      – search engine to analyse the distribution of information throughout time across different media as broadcast television, print media, social media and blogs
      – Mini Research pilot into the relation between media and Eastern European migration
• CLARIAH-SEED: Major Results
  – Demonstrators
    • **OHT and OHT+** (Oral History Today)
      – 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
• CLARIAH-SEED: Major Results
  – Demonstrators
    • **Nederlab** + *De Gids*
      – Search & Analysis interface to historical Dutch text corpora
      – Mini research pilot on author distribution in journal *De Gids*
CLARIAH-SEED: Major Results

- Data

  - **Stakingsdata** (‘Strike Data’)
    - links labour conflicts micro and macro level
    - visualizes them in dynamic historical maps.
  
  - **HLZ (HSN Links Zeeland)**
    - HSN Zeeland and LINKS database combined
  
  - **Athena**
    - Design for historical database on flora and fauna species in cultural and natural contexts for NL
OVERVIEW

• The Humanities are turning Digital
• European Context
• National context

 CLARIAH CORE

• Conclusions
CLARIAH CORE

• 2015-2018
• National project for CLARIN & DARIAH
• Technical Infrastructure
• Linguistics
• Social Economic History
• Media Studies
• Research Pilots
• (Education, dissemination, management)
Technical Infrastructure

- Facilities for shared vocabularies
- Facilities for (mainly meta)data as Linked Data
- Search in the linked data (LD)
- Access Control
- OCR / Text Correction Pipeline
- Standardization
- Performance and Availability
CLARIAH-CORE

• Linguistics
  – Support for the researcher in each stage of a research project
  – New / improved functionality
  • (to the extent that the budget allows it)
  – Cooperation projects with Soc Econ & Media Studies
• Research project stages:
  – Data / Tool creation, collection, crowd-sourcing
  – Browsing / Searching for and selecting data / tools
  – Enrich data
  – Browse and search in (enriched) data + metadata; enrich, analyse, visualise search results
  – Incorporate data / tools in CLARIAH
  – Create enhanced publications
• Social Economic History
  – Databases
    • Macro (national/international)
    • Meso (trade unions, organisations)
    • Micro (individual / family )
  – All structured differently
  – Different vocabularies
  – Integrate these using the Linked Data paradigm
  – To investigate connections / relations
• Media Studies
  – Support the researcher: improved versions of
    • CoMerDA
      – Aggregated Search interface for AV-data
    • AVResearcherXL
      – Exploring audiovisual metadata in historical context
    • TrOve (see above)
• Media Studies
  – Support the researcher: improved versions of
    • DIVE
      – Presentation of collection items in context and ‘intuitive’ browsing
    • OHT (see above)
• Research Pilots
  – Small research projects
  – Aimed at testing infrastructure (elements)
  – Leads to
  • Improved tools and/or
  • Successfully concluded research, and/or
  • New Requirements for tools

• Call will be launched in 2016
• The Humanities are turning Digital
• European Context
• National context
• CLARIAH CORE

ÇConclusions
CONCLUSIONS

• NL is actively participating in CLARIN and DARIAH

• Has a leading role in CLARIN

• Successful in securing funding for RIs

• CLARIN is yielding new research in NL

• Many improvements still needed
  – CLARIAH-CORE just starting
  – Focusing on 3 core disciplines
For more information

www.clariah.nl

THANK YOU FOR YOUR ATTENTION
DO NOT ENTER HERE!
### Filter by research domain:

- **Linguistics (27)**
  - Syntax (13)
  - Morpho-syntactic (7)
  - Historical linguistics (5)
  - Lexicology (5)
  - Dialectology (4)
  - Sign language (4)
  - Acquisition (2)
  - Discourse (2)
  - Language
documentation (2)
- **Morphology (2)**
- **Orthography (2)**
- **Phonetics (2)**
- **Semantics (2)**
- **Onomatology (1)**
- **Phonology (1)**
- **Sociolinguistics (1)**
- **Typology (1)**
- **History (14)**
  - Art History (2)

### Table:

<table>
<thead>
<tr>
<th>TITLE</th>
<th>RESEARCH DOMAIN</th>
<th>TOOL TASK</th>
<th>RESOURCE TAGS</th>
<th>LANGUAGE</th>
<th>CLARIN CENTRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAM-LR</td>
<td>Linguistics, Language Documentation</td>
<td>annotation, audio-visual processing, diarization</td>
<td>service, speech processing, web-service, multi-lingual</td>
<td>Language independent</td>
<td>MPI for Psycholinguistics</td>
</tr>
<tr>
<td>Arthuarian Fiction</td>
<td>Literary Studies</td>
<td>browse, search</td>
<td>data, service, web-application</td>
<td>Huygens ING</td>
<td></td>
</tr>
<tr>
<td>MIMORE</td>
<td>Linguistics, Morpho-syntactic Syntax Dialectology</td>
<td>analysis, browse, search, visualization</td>
<td>service, web-application, mono-lingual</td>
<td>Dutch</td>
<td>Meertens Institute</td>
</tr>
<tr>
<td>TQE</td>
<td>Speech recognition, Linguistics, Phonetics</td>
<td>evaluation</td>
<td>service, speech processing, web-application</td>
<td>Language independent</td>
<td>MPI for Psycholinguistics</td>
</tr>
<tr>
<td>TICClips</td>
<td>Linguistics, Orthography</td>
<td>text processing, orthographic normalisation</td>
<td>service, text processing, web-application</td>
<td>Dutch</td>
<td>Institute for Dutch Lexicology</td>
</tr>
<tr>
<td>INTER-VIEWS</td>
<td>History, Oral History, Linguistics, Discourse</td>
<td>annotation, search</td>
<td>data corpus, multi-media/multi-modal, mono-lingual</td>
<td>Dutch</td>
<td>Meertens Institute</td>
</tr>
</tbody>
</table>
FACETED SEARCH: SYNTAX

CLARIN NL Resource List

Filter by research domain:

- Linguistics
  - Morphology (7)
  - Historical Linguistics (3)
  - Lexicology (3)
  - Morphology (2)
  - Acquisition (1)
  - Dialectology (1)
  - Discourse (1)
  - Orthography (1)
  - Semantics (1)
  - Typology (1)
- Computational Linguistics (2)
- Communication & Media Studies (1)
- History (1)
  - Oral History (1)
- Religion Studies (1)

<table>
<thead>
<tr>
<th>TITLE</th>
<th>RESEARCH DOMAIN</th>
<th>TOOL TASK</th>
<th>RESOURCE TAGS</th>
<th>LANGUAGE</th>
<th>CLARIN CENTRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMORE</td>
<td>Linguistics, Morphology, Syntax, Dialectology</td>
<td>analysis, browse, search, visualisation</td>
<td>service, web-application, mono-lingual</td>
<td>Dutch</td>
<td>Meertens Institute</td>
</tr>
<tr>
<td>DuELME</td>
<td>Linguistics, Lexicology, Syntax</td>
<td>browse, search</td>
<td>data, lexicon, service, web-application, mono-lingual</td>
<td>Dutch</td>
<td>Institute for Dutch Lexicology</td>
</tr>
<tr>
<td>Adelheid</td>
<td>Linguistics, Syntax, Historical Linguistics</td>
<td>annotation, visualisation, text processing, POS tagging, tokenization, lemmatisation</td>
<td>service, web-application, web-service</td>
<td>Dutch</td>
<td>MPI for Psycholinguistics</td>
</tr>
<tr>
<td>TDS</td>
<td>Linguistics, Morphology, Syntax, Typology</td>
<td>browse, search</td>
<td>service, web-application, web-service</td>
<td>DANS</td>
<td></td>
</tr>
<tr>
<td>INPOLDERS</td>
<td>Linguistics, Syntax, Historical Linguistics</td>
<td>parsing</td>
<td>service, text processing, web-application, mono-lingual</td>
<td>DANS</td>
<td></td>
</tr>
</tbody>
</table>
FACETED SEARCH: PARSE
PaQu

PaQu - Parse and Query

Summary
The PaQu web service makes it possible to search in syntactically annotated corpora in Dutch. You can parse your own Dutch text corpus or use one of two corpora provided by the developers.

Background
Please note: This is a preliminary version of this application. Use Lassy (http://dev.clarin.nl/node/1966) for searching the Lassy Treebanks.

PaQu uses the Alpino parser to make treebanks of your own text corpus, and to search in these treebanks using an interface based on the LASSY Word Relations Search interface (http://dev.clarin.nl/node/1966). Two treebanks are already available in the application: Lassy Klein (1M words, manually checked syntactic analysis) and Lassy Groot (700M words, syntactic analysis automatically assigned by Alpino).

PaQu offers two ways to search through the syntactically annotated texts. The first option is to use the search bar to look for word pairs, optionally complemented by their syntactic relationship. The second search option is to use the query language XPath.

Contacts
- Project leader: Drs. P. Kleiweg (Groningen University)
- CLARIN center: none
FACETED SEARCH: ART HISTORY

CLARIN NL Resource List

Filter by research domain:
- History
- Art History
- Cultural Sciences (1)
- Literary Studies (1)

Filter by resource tags:
- service (2)
- data (1)

Filter by tool task:
- browse (2)
- search (2)
RemBench

RemBench - a Digital Workbench for Rembrandt Research

Summary
RemBench enables one to search and browse for works of art, artists, primary sources and library sources related to Rembrandt, using faceted search by location, author/artist name, author/artist type, and date range, and/or by both exact and fuzzy keyword search. It offers both a web application and a RESTful web service.

Background
RemBench combines the content of four different databases behind one search interface:

- RKDartists and RKDimages, two databases maintained by the Netherlands Institute for Art History (RKD)
- RemDoc, a collection of original documents related to Rembrandt van Rijn from the period between 1475 to circa 1750
- RUQuest, a library system that provides access to full text articles, as well as the complete collection of (e-)books and journals from the Radboud University Library Catalogue

RemBench does not influence the content of these databases. For questions or comments about the information in the RKD databases, RemDoc or RUQuest, we refer to their own websites.

Contacts
- Project leader: Suzan Verberne
- CLARIN center: Huigens ING
- Help contact: s.verberne@let.ru.nl

Links
CLIO-INFRA:
HISTORICAL DATA ON WORLD-WIDE
ECONOMIC GROWTH & INEQUALITY

GDP per capita, 1270-2000 ($ of 1990)
Average years of education per capita, 1880 (reconstruction) and 2000
INCOME PER CAPITA AND LIFE EXPECTANCY IN CHINA, CONGO, NL, 1800-2007
Who says what to whom, when, and why?
Which party interrupted which other party and how often?
Who referred to whom and/or what? Named entity detection