# CLARIN - NL

## Language Resources and Technology Infrastructure for the Humanities in the Netherlands

*Jan Odijk*

*NO-CLARIN Meeting*

*Oslo 18 June 2010*

# Overview

- The CLARIN-NL Project
- CLARIN Infrastructure
- Targeted Users
- Subprojects
- Relation with CLARIN-EU
- Future
- Conclusions

# CLARIN-NL Project

- Project in the Netherlands
- Aims to play a leading role in the creation of the European CLARIN technical infrastructure
- Budget: 9.01M Euro
- 2009-2014
- Coordinated by Utrecht University
- 23 participants
- http://www.clarin.nl/

# CLARIN-NL: Infrastructure

- The CLARIN Infrastructure
  - – Will make data and tools on different locations easily accessible via web interfaces and services (CLARIN-portal(s) with intelligent searching, browsing, viewing and querying services)
  - – Will make it possible for non-technical researchers to extract / combine/ enrich data (supported by dissemination and training)
  - – Will make available **interoperable** data and tools based on existing standards and best practices

# CLARIN-NL: For whom?

- For researchers that work with language data
  - Humanities
    - Linguistics (broadly construed)
    - Literary and Theatrical Studies
    - Media en Culture
    - History
    - Political Sciences
    - …

# CLARIN-NL: For whom?

- Current Partners (23)
  - Targeted Users
    - Linguistics (10)
    - Culture (2)
    - Lexicography (2)
    - Social History (4)
    - Literature (2)
  - Technology Providers
    - Language technology (6)
    - Speech technology (2)
  - Data Centres and Service providers
    - Data Centres (5)
    - Libraries (2)

# CLARIN-NL: Subprojects

- Infrastructure Implementation
  - General
    - Partners: Candidate CLARIN Centres
      - MPI, MI, INL, DANS
    - Directly assigned subprojects
    - Provide guidelines / training for others
  - Metadata project (.5 yr)
    - Testing CMDI against existing national data
    - Create initial set of required metadata components

# CLARIN-NL: Subprojects

- Infrastructure Implementation (cont.)
  - Infrastructure Implementation (3 yrs)
    - infrastructure services, an open archiving service, registries, federation of centres, set up a schema registry, profile matching, ISOCAT maintenance, add relation registry RELCAT.
    - coordinate and give guidance for work on web services, wrapper and service bus specification and implementation, select work flow tools and experiment with them.
  - Search&Develop (3 yrs)
    - centralized metadata search
    - distributed content search
      - Text based and structured search

# CLARIN-NL: Subprojects

- User Survey & Base Line (.5yr)
  - Directly assigned subproject
  - User survey
    - Interactive interviews
  - Current use/non-use of digital data and tools
    - Identify causes for non-use
    - Identify obstacles for (wider) use

# CLARIN-NL: Subprojects

- Data Curation & Demonstrator Projects
  - Data Curation project
    - adapt existing resource making it visible, uniquely referable and accessible via the web, and properly documented
  - Demonstrator projects
    - Create a documented web application
      - that can be used as a demonstrator
      - starting from an existing tool or application
      - that can function as a showcase of functionality CLARIN will support

# CLARIN-NL: Subprojects

- Data Curation & Demonstrator Projects
  - Common Goals
    - apply standards and best practices and make use of the suggested CLARIN architecture and agreements
      - esp. CMDI & ISOCAT
      - to understand their limitations and the requirements for extensions
    - establish requirements and desiderata for the CLARIN infrastructure.

# CLARIN-NL: Subprojects

- Data Curation & Demonstrator Projects
  - Must involve a targeted user and address the user’s research questions
  - Open call for subprojects
  - Small subprojects (.5 yr / 60k Euro)
  - 17 projects submitted, 11 received funding
  - Will make available
    - a range of curated resources
    - a range of showcases of CLARIN functionality
    - evidence-based requirements and desiderata
      - for the CLARIN infrastructure and
      - for supported standards and best practices

# CLARIN-NL Subprojects

- INTER-VIEWS project;
  - – Data curation and search functionality for (spoken) interviews with veteran soldiers (Veteraneninstituut)
- AAM-LR
  - – Annotation tool for (field)linguists
  - – mark speech/non-speech
  - – Mark different speakers

# CLARIN-NL Subprojects

- TTNWW (speech)
  - Implement user friendly workflow services for indexing and search of (a limited set of) audio and video data
  - For social historians (Aletta, KDC, KADOC, M2P)
- TICClops (Tilburg)
  - Text cleaning, spelling correction and normalisation

# CLARIN-NL Subprojects

- Adelheid (Nijmegen)
  - Text cleaning, PoS tagging and lemmatisation
  - historical Dutch texts (13th century)
  - For historical linguistic research
- ‘Geleerdenbrievenproject’ (CKCC)
  - selected in the CLARIN-EU call for humanities and social sciences projects as the project proposal
    - that “[would] best demonstrate the use of LRT and would show the potential of a research infrastructure in the humanities”
  - Enriching scholars’ letters with syntactic and semantic annotations
  - In accordance with CLARIN standards
  - For research into circulation of knowledge in scholars’ letters in NL in the 17th century

# CLARIN-NL Subprojects

- (LASSY demo):
  - Simple (‘Google-style’) search interface to automatically parsed text corpora
- TTNWW (text)
  - Implement user friendly workflow services for enriching text corpora with annotations
  - For literature researchers (Huygens) and archeologists (Salagassos)

# CLARIN-NL Subprojects

- Standardisation and integration of linguistic data and tools (for linguistic research)
  - En Garde/DUELME-LMF (UU)
    - DUELME database of multi-word expressions
  - WFT-GTB (Fryske Akademy)
    - Integration of *Wurdboek fan 'e Fryske Taal* with Integrated Language Data Base
  - ADEPT (UG)
    - Adaptation of edit-distance tool for dialect and historical linguistic research

# CLARIN-NL Subprojects

- Standardisation and integration of linguistic data and tools (for linguistic research)
  - MIMORE (MI, UU)
    - Microcomparative Morphosyntax Research Tool
  - TDS-Curator (UU)
    - Curation of the Typological Database System
  - TQE (RU)
    - Transcription Quality Evaluation
  - Sign-LinC (RU)
    - Links lexical databases and annotated corpora of sign languages

# CLARIN-NL: Subprojects

- Education, Training & Awareness
  - Organize conferences / workshops / meetings
  - Attend / presentations at events
  - Support events (logistically and financially)
  - Support individual researchers for visiting events
  - Tutorials and lectures (ISOCAT, CMDI, PIDs, …), presence at Summer and Winter schools
  - Website (with Web2.0 functionality)
  - Newsflashes, newsletters, etc.

# CLARIN-NL: v. CLARIN-EU

A Pan European Initiative

- Organizationally
  - A CLARIN ERIC is being set up
  - NL aims to host the CLARIN ERIC
  - Dutch Minister of Education, Culture and Sciences invited his colleagues to join the CLARIN ERIC
  - CLARIN-NL has funds to fulfil a leading role of NL in the CLARIN ERIC

# CLARIN-NL: v. CLARIN-EU

- Content-wise
  - Complementary to EU preparatory project
    - Not only preparatory phase but also implementation phase and first part of exploitation phase
    - carries out activities such as the metadata project, the data curation and demonstrator projects
      - Focusing on data and tools from the Netherlands
      - Not covered by the European preparatory project

# CLARIN-NL: Future

- Working on priorities for next subprojects (2010-2011)
  - Analyzing current situation, identifying gaps
  - Proposals are being worked out on
    - form (open call, tender, direct assignment, mix)
    - focus on topics/disciplines
    - Budget and timing
  - Decisions expected Mid 2010
- Centres of Expertise
  - Centres of expertise are physical or virtual centres that possess a specific type of knowledge and expertise on a topic that is relevant to CLARIN and that have sufficient mass to guarantee the sustainability of this knowledge and expertise.
  - Identify candidates
  - Plan of activities

# CLARIN-NL: Future

- Embedding
  - start work for ensuring the longer term existence of the CLARIN infrastructure
  - embed it in the normal research activities
  - prepare both a governance structure and structural financing
  - in close cooperation with CLARIN EU
- Continue and intensify educational, training and awareness activities
  - In particular get the CLARIN infrastructure and working in a CLARIN-compatible manner into the regular curricula of universities.

# CLARIN-NL: Conclusions

- **the CLARIN-NL project is an excellent example for other national CLARIN projects**
  - mix between a programme and a project
    - provides flexibility (new developments, new players)
    - offers opportunities for defining a few longer term projects in selected areas to sustain knowledge and expertise built up in the participating institutes.
  - data curation and demonstrator projects
    - offer opportunities for testing standards and best practices and CLARIN architecture
    - will strengthen these and also show their limitations
    - Will provide evidence-based arguments for modifications or extensions
    - Provides opportunities for influencing a selection of standards and best practices compatible with the existing national data.
    - will yield curated data, a range of showcases for explaining and demonstrating the advantages of the CLARIN infrastructure and the new possibilities it will offer

# CLARIN-NL: Conclusions

– Involvement of targeted users

- cooperation between the targeted users and the technology providers is required
- with a central role for the users' research questions
- bringing these different communities together in concrete cooperation projects
- So that the CLARIN infrastructure will provide the functionality that is actually needed by the researchers

# CLARIN-NL

Thanks for your attention!

http://www.clarin.nl/

# CLARIN-NL Governance

- – Executive Board (4)
- – National Advisory Panel (17)
- – International Advisory Panel (6)
- – Board (8)