CLARIN for Linguists
Storing resources in CLARIN

Jan Odijk
LOT Summerschool
Nijmegen, 2014-06-27
Overview

• Why store resource in CLARIN?
• How to store resources in CLARIN
  – What you must do
  – What the CLARIN Centre must do
Why store resource in CLARIN?

• How to store resources in CLARIN
  – What you must do
  – What the CLARIN Centre must do
Why?

• You may benefit from it
  – Existing tools in CLARIN
    • faster production, better quality, more features
    • Search engines, analysis tools, visualisation tools
  – Can be easily combined with other data in CLARIN

• Others may benefit from it
  – Many unexpected uses of your data
  – Now or in the future
Why?

- Openness in Science
  - Usually produced with public money
- Integrity
  - Recently many scandals with faked data
- Verifiability and Replicability of research results
  - Essential for the proper conduct of science
  - More and more journals are requiring it
- Funding Agency requires it (data management plan)
- DANS CLIP on data sharing (in Dutch)
• Why store resource in CLARIN?
  ➢ How to store resources in CLARIN
    – What you must do
    – What the CLARIN Centre must do
• Start early
  – Preferably before you start creating data
• Contact a CLARIN Type B centre
  – They can help you
  – Your resource must be stored at a CLARIN centre
• Which CLARIN centre?
  – Check CLARIN-NL portal (http://dev.clarin.nl)
  – CLARIN-NL website Centres page
  – Brief summary
• Why store resource in CLARIN?
• How to store resources in CLARIN
  ➢ What you must do
    – What the CLARIN Centre must do
What you must do

• Define what your data are / are going to be
• Ensure legal / ethical compliance
  – Permission from subjects to use the data for research
  – Provisions for respecting privacy matters
• Determine Metadata Contents
  – Determine what information should be included in the metadata (resource description) of your resource
  – Collect this information
WHAT YOU MUST DO

• Determine centre
• **Contact them** and make arrangements
• Determine **CLARIN-recommended standard format(s) for your data**
  – Consult with the CLARIN Centre
  – Ask help from the helpdesk [helpdesk@clarin.nl](mailto:helpdesk@clarin.nl)
  – E.g. LMF (Lexical Markup Framework for lexicons, cf. Cornetto, DuELME, ...)
  –
What you must do

- Metadata must be in CMDI format
- CMDI provides
  - A model for metadata
  - A format for metadata
  - Tools to make metadata
- CMDI Metadata are written in XML
- It does NOT proscribe the contents of the metadata
- [Introduction to CMDI](mailto:cmdi@clarin.eu)
What you must do

- CMDI Metadata use a metadata profile
- **Metadata profile**: a combination of
  - Metadata components
  - Metadata elements
- **Metadata Component**: a combination of
  - Metadata components (optional)
  - Metadata elements (optional)
- **Metadata Element**
  - XML element: name, value (of an explicit type), attribute-value pairs
• This provides high flexibility
  – YOU can determine the metadata for your resource
  – By defining your own profiles, components, elements
• CMDI helps you with
  – A profile and component editor  [login required]
  – A list of commonly used profiles and components
  – A metadata editor: ARBIL
• Flexibility requires explicit semantics!
  – The CLARIN infrastructure must `know’ what you mean with your metadata elements
  – Otherwise it cannot use faceted browsing in the VLO or the Meertens Metadata Search Engine
Explicit semantics:

- Each element in the **data and metadata** must have a link to a CLARIN-recognized concept or data category registry
- Most prominent data category registry in CLARIN is **ISOCAT**
- [Example Data Category in ISOCAT](#)
- [Example Link to ISOCAT in element definition](#)
Explicit semantics (2):

- **RELCAT alpha version**
  - For relations between data categories

- **SCHEMACAT alpha version**
  - For describing schemas of resources

- **CLAVAS Vocabulary Service**
  - Interface to other data category registries and vocabularies
    - **ISO 639-3 Language codes**
What you must do

• Attend dedicated tutorials on CMDI and ISOCAT
  – Regularly organized in NL (each 2+ times/year)
• Usually the CLARIN Centre helps you creating the CMDI metadata
• Maximally reuse existing profiles /components
  – It will help you get better metadata
  – You do not have to reinvent the wheel
CLARIN strongly recommends using certain components (e.g. GeneralInfo component) and may require inclusion of certain properties.

Do not forget properties that are ‘obvious to you’, e.g.
- Language
- Title
- Version
What you must do

• Live Version v. exchange/archive version
  – E.g. lexicon
    • in Lexical Markup Framework compatible XML text v.
    • database with indexes for fast search
  – Live version is ideally derived fully automatically from the exchange/archive version
• In close cooperation with the CLARIN Centre
What you must do

- Software
  - Desktop tool
  - Desktop application
  - Web service
    - SOAP / REST, use CLAM if possible
  - Web application
- There must be `metadata' for software as well!
  - Generic profile exists and is being refined
- Consult with the CLARIN Centre
• Why store resource in CLARIN?
• How to store resources in CLARIN
  – What you must do
  ➢ What the CLARIN Centre must do
• Assist you with your tasks
• Assign Persistent Identifiers (PIDs) to all data and metadata
  – Handle system for assignment and resolution of PIDs
  – example
• Make Metadata harvestable
  – OAI-PMH protocol
    • Open Archives Initiative Protocol for Metadata Harvesting
• Store the data in the centre’s repository
  – LAMUS (the Language Archive) and its documentation online or as PDF
  – EASY (DANS) and its Help and Support Page
• Make data themselves available and accessible in the CLARIN infrastructure
What the CLARIN Centre must do

- Provisions for legal / ethical restrictions
- Long term preservation
  - Data Seal of Approval
- (Minimal) Maintenance
What about existing data?

- Adapt them to meet the CLARIN requirements
  - Data Curation
- CLARIN-NL
  - Has financed many data curation projects
  - Has set up Data Curation Service
- CLARIAH (successor project, if financed)
  - Will continue these curation activities
Thanks for your attention!
• **Meertens Institute**: resources relevant for the study of
  – cultural expressions and Language variation within the Dutch language

• **Max Planck Institute for Psycholinguistics (The Language Archive)**: resources related to the study of
  – psychological, social and biological foundations of language

• **Huygens Institute**: resources related to the study of
  – history and literature of the Netherlands.

• **Institute for Dutch Lexicology (INL)**
  – relevant to the lexicological study of the Dutch language

• **Data Archiving and Networked Centres (DANS)**
  – digital research data generally
CLARIN-NL Centres

• Koninklijke Bibliotheek (National Library):
  – Digital books, articles, newspapers
  – Includes DBNL (Digital Library for Dutch Literature)
  – (will be available in the VLO soon)

• Nederlands Instituut voor Beeld & Geluid (NIBG, Netherlands Institute for Sound and Vision )
  – Audio-visual data (esp. TV and radio programmes)
  – NIBG data via the VLO

• Utrecht University Library (UBU)
  – Digital books, articles
  – UBU data via the VLO
### CMDI Profile: Example

#### Component Registry

<table>
<thead>
<tr>
<th>Name</th>
<th>Group Name</th>
<th>Domain Name</th>
<th>Creator</th>
<th>Description</th>
<th>Registration Date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTRP_sub_location</td>
<td></td>
<td></td>
<td>Folkert de Vriend</td>
<td>Profile for locations subdivision</td>
<td>30 July 2010</td>
<td>0</td>
</tr>
<tr>
<td>imdr-corpus</td>
<td>Alexander Koeng</td>
<td></td>
<td></td>
<td>IMDI corpus profile</td>
<td>02 June 2010</td>
<td>0</td>
</tr>
<tr>
<td>imdr-session</td>
<td>CLARIN (<a href="mailto:cmdi@clarin.eu">cmdi@clarin.eu</a>)</td>
<td></td>
<td></td>
<td>IMDI metadata (type: session)</td>
<td>21 April 2010</td>
<td>3</td>
</tr>
<tr>
<td>LexicalResourceProfile</td>
<td>Laura van Eerten</td>
<td></td>
<td></td>
<td>a profile for describing a lexical</td>
<td>26 April 2010</td>
<td>2</td>
</tr>
</tbody>
</table>

#### LexicalResourceProfile

- **Name**: LexicalResource
- **Group name**: clarin-nl/lexical/resource
- **Description**: A component for describing lexical resources like different types of word lists, dictionaries and semantic networks.

**Components**

- **cmdi-generalinfo**
  - Number of occurrences: 1 - 1

- **OriginLocation**
  - Number of occurrences: 0 - unbounded
    - **cmdi-location**
      - Number of occurrences: 1 - 1

- **cmdi-project**
  - Number of occurrences: 0 - unbounded

- **cmdi-creators**
  - Number of occurrences: 0 - 1

- **cmdi-importance**
  - Number of occurrences: 0 - 1
CMDI Component: Example

<table>
<thead>
<tr>
<th>Name</th>
<th>Group Name</th>
<th>Domain Name</th>
<th>Creator</th>
<th>Description</th>
<th>Registration Date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>cmdi-language</td>
<td>clarin</td>
<td>CLARIN (<a href="mailto:cmdi@clarin.eu">cmdi@clarin.eu</a>)</td>
<td>Component for describing a...</td>
<td>21 April 2010 16:17:21</td>
<td>21 April 2010 16:17:21</td>
<td>0</td>
</tr>
<tr>
<td>cmdi-location</td>
<td>clarin</td>
<td>CLARIN (<a href="mailto:cmdi@clarin.eu">cmdi@clarin.eu</a>)</td>
<td>Component for describing a...</td>
<td>21 April 2010 16:17:21</td>
<td>21 April 2010 16:17:21</td>
<td>0</td>
</tr>
<tr>
<td>cmdi-mimetype</td>
<td>clarin</td>
<td>CLARIN (<a href="mailto:cmdi@clarin.eu">cmdi@clarin.eu</a>)</td>
<td>List of frequently used mime...</td>
<td>31 January 2013 12:48:14</td>
<td>31 January 2013 12:48:14</td>
<td>0</td>
</tr>
<tr>
<td>Creation</td>
<td>CLARIN</td>
<td>CLARIN</td>
<td>halida</td>
<td>Component contains informati...</td>
<td>31 January 2013 12:48:14</td>
<td>0</td>
</tr>
</tbody>
</table>

**cmdi-location**

<table>
<thead>
<tr>
<th>Name: Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Name: clarin</td>
</tr>
<tr>
<td>Description: Component for describing a certain location (address, region, country, continent)</td>
</tr>
</tbody>
</table>

**Elements**

- **Address string**
  - ConceptLink: http://www.isocat.org/dataset/DC-2528
  - Number of occurrences: 0 - 1

- **Region string**
  - ConceptLink: http://www.isocat.org/dataset/DC-2533
  - Number of occurrences: 0 - 1

**Component**

- **iso-country**
  - Number of occurrences: 0 - 1

- **iso-continent**
  - Number of occurrences: 0 - 1
CMDI Element: Example

### Component Registry

**Component Browser**

<table>
<thead>
<tr>
<th>Name</th>
<th>Group Name</th>
<th>Domain Name</th>
<th>Creator</th>
<th>Description</th>
<th>Registration Date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>media-session</td>
<td>CLARIN-D Version 1.2</td>
<td>phonetics</td>
<td><a href="mailto:71A0D0C90E89334@imu.de">71A0D0C90E89334@imu.de</a></td>
<td>Group of recordings of at least 30 minutes.</td>
<td>05 February 2013 14:12:47</td>
<td>0</td>
</tr>
<tr>
<td>WrittenResource</td>
<td>CLARIN-D Version 1.2</td>
<td>phonetics</td>
<td><a href="mailto:71A0D0C90E89334@imu.de">71A0D0C90E89334@imu.de</a></td>
<td>Component contains information about a textual resource.</td>
<td>05 February 2013 14:13:45</td>
<td>0</td>
</tr>
<tr>
<td>author</td>
<td>CLARIN-D: DTA-Basisformat</td>
<td></td>
<td>Axel Harold</td>
<td>the author of a textual resource.</td>
<td>16 September 2013 15:59:33</td>
<td>0</td>
</tr>
</tbody>
</table>

**fileDesc**

- **Element:** title
  - **ConceptLink:** http://www.isocat.org/dcat/DC-2545
  - **Documentation:** the title of a work according to the classification in cotype
  - **DisplayPriority:** 1
  - **Number of occurrences:** 1 - unbounded
  - **Multilingual:** false

- **AttributeList:**
  - **type** : main
  - **level** : a

- **Component:** author
  - **Number of occurrences:** 0 - unbounded

- **Component:** editor
  - **Number of occurrences:** 0 - unbounded
LINK TO ISOCAT DC
• CLARIN attempts to maximize open and free access to resources
  – with as little restrictions as possible
  – no login unless it cannot be avoided
• Sometimes, a login (Authentication and Authorisation, AAI) is required, e.g.
  – Because there are legal and/or ethical restrictions on the data
  – To identify you and assign you your own workspace / data
  – To enter with your own personal settings
CLARIN is a distributed infrastructure

- How can we avoid that you have to login again and again?
- How can we avoid that you have to remember many user names and passwords?
- How can we avoid that CLARIN has to securely store user names, passwords and possibly other privacy-sensitive information?
The answer: **Shibboleth**

- When you log in, you are directed to a login with your own institute.
- You then log in with your institute’s user name and password.
- The institute server then confirms that you are a trusted person, and you can enter this part of the CLARIN infrastructure.
  - It does **not** pass on any sensitive information such as your user name or password.
- If you now go to another part of the CLARIN infrastructure that requires login, it ‘knows’ that you are already logged in, so you do not have to do this again (Single Sign On, SSO).
• Return Page
## Component Browser

<table>
<thead>
<tr>
<th>Name</th>
<th>Group Name</th>
<th>Domain Name</th>
<th>Creator</th>
<th>Description</th>
<th>Registration Date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnnotationTool</td>
<td></td>
<td></td>
<td>Eric Sanders</td>
<td>Description of a tool adapted from...</td>
<td>09 February 2011 14:13:59</td>
<td>0</td>
</tr>
<tr>
<td>ArthurianFiction</td>
<td>Other</td>
<td></td>
<td>Rik Hooistra</td>
<td>Profile for Arthurian Fiction data...</td>
<td>04 September 2012 14:20:05</td>
<td>0</td>
</tr>
<tr>
<td>Bamdo.textualResource</td>
<td>computational_linguistics</td>
<td></td>
<td>Dieter Van Uytvanck</td>
<td>Lexical Resource as used by BAMBES...</td>
<td>27 October 2010 15:47:42</td>
<td>0</td>
</tr>
<tr>
<td>BamdoModalCorpus</td>
<td>computational_linguistics</td>
<td></td>
<td>Dieter Van Uytvanck</td>
<td>Oral Corpus as used by BAMBES...</td>
<td>27 October 2010 16:00:41</td>
<td>0</td>
</tr>
<tr>
<td>BamdoOralCorpus</td>
<td>computational_linguistics</td>
<td></td>
<td>Dieter Van Uytvanck</td>
<td>Oral Corpus as used by BAMBES...</td>
<td>27 October 2010 16:01:05</td>
<td>0</td>
</tr>
<tr>
<td>BamdoTool</td>
<td>computational_linguistics</td>
<td></td>
<td>Dieter Van Uytvanck</td>
<td>Tool as used by BAMBES (for th...</td>
<td>27 October 2010 15:48:54</td>
<td>0</td>
</tr>
<tr>
<td>BamdoWrittenCorpus</td>
<td>computational_linguistics</td>
<td></td>
<td>Dieter Van Uytvanck</td>
<td>Written Corpus as used by BAMBES...</td>
<td>27 October 2010 15:59:28</td>
<td>0</td>
</tr>
<tr>
<td>Boedelbank</td>
<td></td>
<td></td>
<td>Folkart de Vriend</td>
<td>Profile for Boedelbank</td>
<td>20 July 2010 9:33:18</td>
<td>0</td>
</tr>
<tr>
<td>Boedelbank</td>
<td></td>
<td></td>
<td>Folkart de Vriend</td>
<td>Profile for Boedelbank</td>
<td>30 July 2010 8:36:25</td>
<td>0</td>
</tr>
<tr>
<td>cmdl-virtual-collection</td>
<td></td>
<td></td>
<td>CLARIN (<a href="mailto:cmdl@clarin.eu">cmdl@clarin.eu</a>)</td>
<td>A component for describing per...</td>
<td>21 April 2010 16:18:30</td>
<td>0</td>
</tr>
<tr>
<td>collection</td>
<td></td>
<td></td>
<td>Matej Durco</td>
<td>minimal collection profile (start...</td>
<td>15 October 2010 20:38:27</td>
<td>0</td>
</tr>
</tbody>
</table>
Adelheid

This webservice annotates historical Dutch texts with word class tags and lemmas.

Start a new Project

Project ID: [ ] Start project

Projects

Search: [ ] Last changed

No data available in table

Powered by CLAM v0.7.6 - Computational Linguistics Application Mediator

Induction of Linguistic Knowledge Research Group, Tilburg University

CLAM is funded under CLARIN-NL projects EUOZCOPS (09-011) coordinated by Martin Reynaert, and TITWIK, WP1 and WP2, respectively coordinated by Martin Reynaert and Annal van den Bosch.
• Return Page