Metadata for tools:
creating a CMDI profile for tools

Eline Westerhout and Jan Odijk

Utrecht University
1. Background

2. The metadata profile
   - Retrievability component
   - Technical component
   - Implementation

3. Conclusions
2009-2014

- projects
  - demonstration (Adelheid, WAHSP, DISCAN)
  - curation (e.g. INTER-VIEWs, C-DSD, WFT-GTB)
  - other (infrastructure implementation, metadata, S&D)
CLARIN-NL

- 2009-2014 projects
  - demonstration (Adelheid, WAHSP, DISCAN)
  - curation (e.g. INTER-VIEWs, C-DSD, WFT-GTB)
  - other (infrastructure implementation, metadata, S&D)

“... aims to design, construct, validate, and exploit a research infrastructure in the Netherlands that is needed to provide a sustainable and persistent eScience working environment for researchers in the Humanities, and, in particular, in Linguistics, who want to make use of language resources and the technology to use these resources for their research.”
Metadata for tools in CLARIN-NL

- current status
  - own implementations
  - inconsistent and incomplete

- need to enhance
  - retrievability
  - consistency
  - completeness
Solution: metadata for tools project

- 3 PM (August 2012-January 2013)
- aim: develop generic metadata profile for tools
- result: profile with two components
  1. retrievability
  2. technical
The retrievability component

- User-oriented
- Description of the software
- No technical details
Virtual Language Observatory

The metadata profile

Retrievability component

Metadata for tools

18 January 2013 7 / 21
Adelheid

A Distributed Lemmatizer for Historical Dutch, version 1.0

This is the official home page of the Adelheid tagger-lemmatizer system. It was developed originally by Hans van Halteren on the basis on the POS-tagged corpus van Reenen-Mulder. Subsequently it was made publicly available through the Clarin infrastructure during the NWO-sponsored Clarin-NL project Adelheid.

On this page you will find links to the system, useful files, and news about the status of the system.

The Adelheid System

Here are links to the system components. Please read the manual (see links below) before use.

- Adelheid Tagger-Lemmatizer
- Adelheid Visualisation Tool

Downloadables

And here are links to manual and example files.

- Tagger-Lemmatizer Manual
- Tagger Manual
- Annotation Tool Manual
- Demonstration Scenarios
- Example files
- DTDs for input and output files

Status

Adelheid has been officially released (Version 1.0).

However, we are currently experiencing the following disturbances:

- If you want to access the systems as web services, you need to apply to hvh@let.ru.nl for a user account. For access as a web application, you can use your institutional login (see manual).
- The Visualisation Tool is not available due to technical problems. This means that you cannot fully go through the demonstration scenarios: instead of viewing the output with the Visualisation Tool, you will have to examine the .tag output with a text editor.
Example: Adelheid

**Adelheid**

A Distributed Lemmatizer for Historical Dutch version 1.0

This is the official home page of the Adelheid tagger-lemmatizer system. It was developed originally by Hans van Halteren on the basis on the POS-tagged corpus of Reenen-Mulder. Subsequently it was made publicly available through the Clarin infrastructure during the NWO-sponsored Clarin-NL project Adelheid.

On this page you will find links to the system, useful files, and news about the status of the system.

**The Adelheid System**

Here are links to the system components. Please read the manual (see links below) before use.

- Adelheid Tagger-Lemmatizer
- Adelheid Visualisation Tool

**Downloadables**

And here are links to manual and example files.

- Tagger-Lemmatizer Manual
- Tagger Manual
- Annotation Tool Manual
- Demonstration Scenarios
- Example files
- DTDs for input and output files

**Status**

Adelheid has been officially released (Version 1.0).

However, we are currently experiencing the following disturbances:

- If you want to access the systems as web services, you need to apply to hvh@let.ru.nl for a user account; for access as a web application, you can use your institutional login (see manual).
- The Visualisation Tool is not available due to technical problems. This means that you cannot fully go through the demonstration scenarios: instead of viewing the output with the Visualisation Tool, you will have to examine the .tag output with a text editor.
The technical component

- Developer-oriented
- Technical details of the software
- Consistent & complete description
Implementation process

1. Identify concepts & design structure
2. Build components
3. Link to ISOcat
4. Build profile
5. Fill in metadata

- Component Registry
- ISOcat & Component Registry
- Component Registry
- Arbil editor
1. Identify concepts & structure

input:
- websites
- metadata
- manuals/reports

structure:
- GeneralInfo
- SoftwareFunction
- SoftwareImplementation
- Access
- Documentation
- SoftwareDevelopment
- TechnicalInfo
- Service
2. Build components

![Component Browser](image)

**Clarin Component Registry**

### Component Browser

<table>
<thead>
<tr>
<th>Name</th>
<th>Group Name</th>
<th>Domain</th>
<th>Creator</th>
<th>Description</th>
<th>Registration Date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Component contains information about the availability and accessibility of a resource.</td>
<td>06 December 2012 11:21:41</td>
<td>0</td>
</tr>
<tr>
<td>AlternativeName</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Alternative name for a resource, e.g. an English name.</td>
<td>05 October 2012 14:56:53</td>
<td>0</td>
</tr>
<tr>
<td>Contact</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Component contains contact information about creators, distributors, actors... of a resource.</td>
<td>15 October 2012 14:46:23</td>
<td>0</td>
</tr>
<tr>
<td>Copyright</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Description of the copyright on a resource.</td>
<td>04 October 2012 15:17:01</td>
<td>0</td>
</tr>
<tr>
<td>CopyrightHolder</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Information on the copyright holder of a resource.</td>
<td>06 December 2012 15:46:56</td>
<td>0</td>
</tr>
<tr>
<td>Creator</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Component contains (contact) information about a person and/or his/her organisation.</td>
<td>15 October 2012 14:38:10</td>
<td>0</td>
</tr>
<tr>
<td>Documentation</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Documentation of a resource.</td>
<td>25 October 2012 12:01:57</td>
<td>0</td>
</tr>
<tr>
<td>GeneralInfo</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Component contains general information about the resource. Extension of cmdl-geninfo.</td>
<td>22 November 2012 14:17:04</td>
<td>0</td>
</tr>
<tr>
<td>ImplementationLanguage</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Indication of the programming language used for implementation, the version of the language.</td>
<td>22 August 2012 14:53:45</td>
<td>0</td>
</tr>
<tr>
<td>Input</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Description of the input characteristics.</td>
<td>27 September 2012 17:03:37</td>
<td>0</td>
</tr>
<tr>
<td>InstallationRequirements</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Software and hardware requirements for the installation of a resource.</td>
<td>05 October 2012 14:36:05</td>
<td>0</td>
</tr>
<tr>
<td>ISO4217</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Standard currency codes referred to as the ISO 4217 code list. Currency codes are used...</td>
<td>11 October 2012 17:26:07</td>
<td>0</td>
</tr>
<tr>
<td>LanguageVariety</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>The component specifies the name of the language variety and the time period(s) for...</td>
<td>22 November 2012 17:36:50</td>
<td>0</td>
</tr>
<tr>
<td>LinguisticsSubject</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>Description of the subject matter of a resource as relevant to a particular subfield of...</td>
<td>27 September 2012 16:57:03</td>
<td>0</td>
</tr>
<tr>
<td>MinimumHardwareConfiguration</td>
<td>CLARIN-NL</td>
<td></td>
<td>Eline Westerhout</td>
<td>The minimum hardware requirements for using a resource.</td>
<td>23 August 2012 15:04:08</td>
<td>0</td>
</tr>
</tbody>
</table>
Example: SoftwareFunction

Name: SoftwareFunction
Group Name: CLARIN-NL
Description: Component contains general information about the function of the software

Element: toolCategory  
- annotationTool
  - Documentation: Corresponding DFKI category of the resource (http://registry.dfk.de/)
  - DisplayPriority: 1
  - Number of occurrences: 1 - unbounded

Element: toolTask  
- abbreviationExpansion
  - ConceptLink: http://www.isocat.org/dcat/DC-2500
  - Documentation: The task(s) carried out by the software or a typical task description of the software.
  - DisplayPriority: 2
  - Number of occurrences: 1 - unbounded

Component: LinguisticsSubject
- Number of occurrences: 1 - unbounded

Component: LanguageVariety
- Number of occurrences: 1 - unbounded

Component: SoftwarePerformance
- Number of occurrences: 0 - unbounded

Component: cmdi-description
- Number of occurrences: 0 - 1
### 3. Link to ISOcat

#### ISOcat Interface

![ISOcat Interface](image)

#### Adelheid Attributes Table

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Version</th>
<th>Administration status</th>
<th>Registration status</th>
<th>Check</th>
<th>Type</th>
<th>Owned by</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>4212</td>
<td>annotator character</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4213</td>
<td>annotator form</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4216</td>
<td>annotator separator</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4218</td>
<td>confidence</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4219</td>
<td>form</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4220</td>
<td>frequency</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4221</td>
<td>lemma</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4222</td>
<td>manuscript character</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4223</td>
<td>manuscript form</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4224</td>
<td>manuscript separator</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
<tr>
<td>4214</td>
<td>potential lemma</td>
<td>1:0</td>
<td>private</td>
<td>private</td>
<td>✓</td>
<td>open</td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
</tbody>
</table>

#### Annotator Character - 1:0

<table>
<thead>
<tr>
<th>Key</th>
<th>PID</th>
<th>Type</th>
<th>Owner</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4212</td>
<td><a href="http://www.isocat.org/dcat/DC-4212">http://www.isocat.org/dcat/DC-4212</a></td>
<td>van Halteren, Hans</td>
<td>private</td>
</tr>
</tbody>
</table>

#### Administration Information Section

1. Administration Record

- Identifier: AChar
- Version: 1:0
- Registration Status: private
- Administration Status: private
- Justification: Attribute used by Adelheid in element charpos
4. Build profile

Name: ClarinSoftwareDescription
Group Name: CLARIN-NL
Description: Description of a software program.

Component: GeneralInfo
Number of occurrences: 1 - 1

Component: SoftwareFunction
Number of occurrences: 1 - 1

Component: SoftwareImplementation
Number of occurrences: 1 - 1

Component: Access
Number of occurrences: 1 - 1

Component: ResourceDocumentation
Number of occurrences: 1 - 1

Component: SoftwareDevelopment
Number of occurrences: 1 - 1

Component: TechnicalInfo
Number of occurrences: 0 - 1

Component: Service
Number of occurrences: 0 - unbounded
5. Fill in metadata - I
5. Fill in metadata - II

```xml
<?xml version="1.0" encoding="UTF-8"?>
<CMD xmlns="http://www.clarin.eu/cmd/
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
CMDVersion="1.1"
ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1342181139640/xsd">
  <Header>
    <MdCreator>elinewesterhout</MdCreator>
    <MdCreationDate>2013-01-11T01:00</MdCreationDate>
    <MdProfile>clarin.eu:cr1:p_1342181139640</MdProfile>
  </Header>
  <Resources>∞</Resources>
  <Components>
    <ClarinSoftwareDescription>
      <GeneralInfo>
        <name>Adelheid tagger-lemmatizer</name>
        <url>http://adelheid.ruhosting.nl</url>
        <ReleaseStatus>
          <LifeCycleStatus>released</LifeCycleStatus>
          <lastUpdate/>
        </ReleaseStatus>
      </GeneralInfo>
      <SoftwareFunction>
        <toolCategory>writtenLanguageTool</toolCategory>
        <toolTask>lemmatization</toolTask>
        <LinguisticsSubject>
          <LinguisticsSubject>historicalLinguistics</LinguisticsSubject>
          <LinguisticsSubject>
        </LinguisticsSubject>
        <LanguageVariety>
          <century>14</century>
          <languageDependent>1</languageDependent>
          <ISO639>
            <iso-639-3-code>nld</iso-639-3-code>
            <ISO639>
          </ISO639>
        </LanguageVariety>
      </SoftwareFunction>
    </ClarinSoftwareDescription>
  </Components>
</CMD>
```
Conclusions

- Metadata creation supported
- Components almost finished
- Advantages
  - software retrievable
  - complete descriptions
Future work

- Finalize the profile
- Fill in the profile
  - completed projects
  - new/running projects
- Create web form
Thank you

Website CLARIN-NL
http://www.clarin.nl

Website CMDI
http://www.clarin.eu/cmdi

Contact
E.N.Westerhout@uu.nl