Metadata for tools: creating a CMDI profile for tools

Eline Westerhout and Jan Odijk

Utrecht University

Outline

1. Background
2. The metadata profile
   - Retrievability component
   - Technical component
   - Implementation
3. Conclusions

CLARIN-NL

- 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

VLO Home >> Faceted Browser Resources search

COLLECTION

Meertens collection: Liederenbank (243129)
Nederlands Instituut voor Beeld en Geluid Academia collectie (46156)
childes (28595)
Endangered Languages (20896)
Language and Cognition (20457)
DK-CLARIN Repository (14427)
talkbank (14243)
Acquisition (13142)
MPI CGN (12769)
Bavarian Archive for Speech Signals (BAS) (11562)
more...

LANGUAGE

Dutch (136452)
Nederlands (109557)
English (67745)
German (29217)
French (11581)
Spanish; Castilian (11189)
Japanese (7012)
Turkish (6179)
Swedish (6065)
Chinese (2173)
more...

CONTINENT

Europe (61531)
North-America (21646)
Asia (16946)
South-America (8051)
Oceania (5006)
Africa (4236)
Australia (2218)

GENRE

discourse (80057)
kinderlied (18594)
primary_text (7350)
liefdeslied (5812)
bespiegelend lied (4722)
language_description (4327)
story telling (3996)

Showing 1 to 10 of 493914 << < *1* 2 3 4 5 6 7 8 9 10 > >>

name	description
!O!ung: a language of Angola	A page from the Web edition of Ethnologue: Languages of the World (16th edition) giving basic facts...
!Xóõ: a language of Botswana	A page from the Web edition of Ethnologue: Languages of the World (16th edition) giving basic facts...
" DE STIEFMOEDER " die haar 18 jarigen STIEFZOON 10 jaar heeft opgesloten op de 3e verdieping Over-Amstelstraat 10 te AMSTERDAM	
"'T IS FEEST."	
"ACH, MIJN LIEVE AUGUSTIJN."	

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

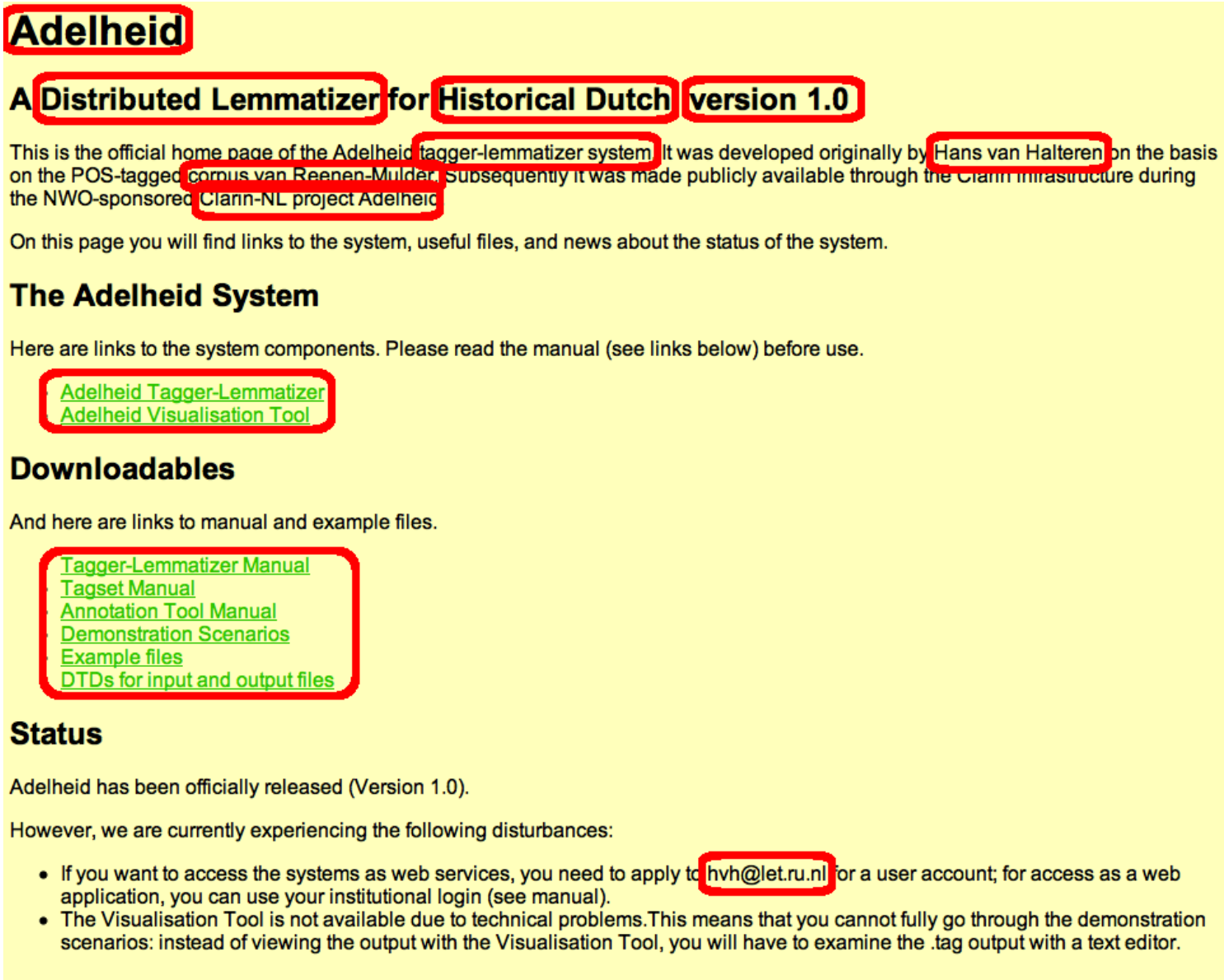
Status

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

Status

Adelheid has been officially released (Version 1.0).

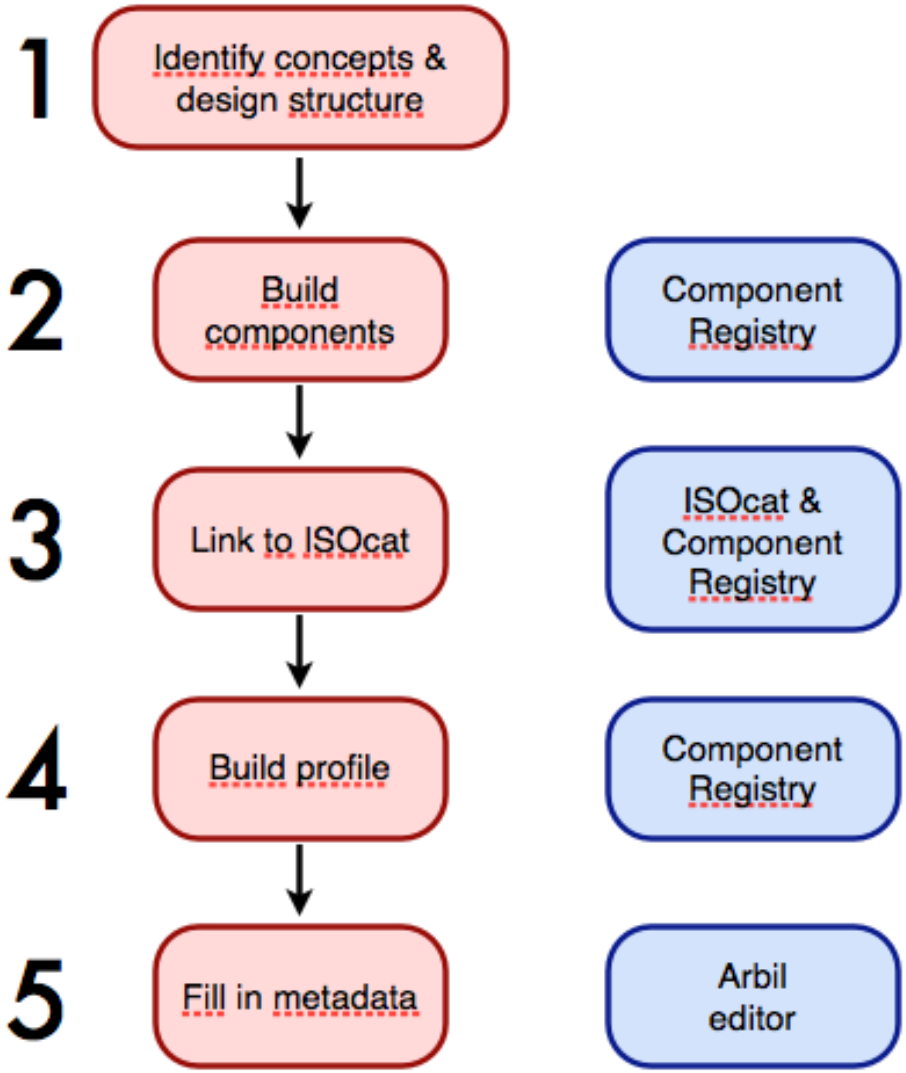
However, we are currently experiencing the following disturbances:

- 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

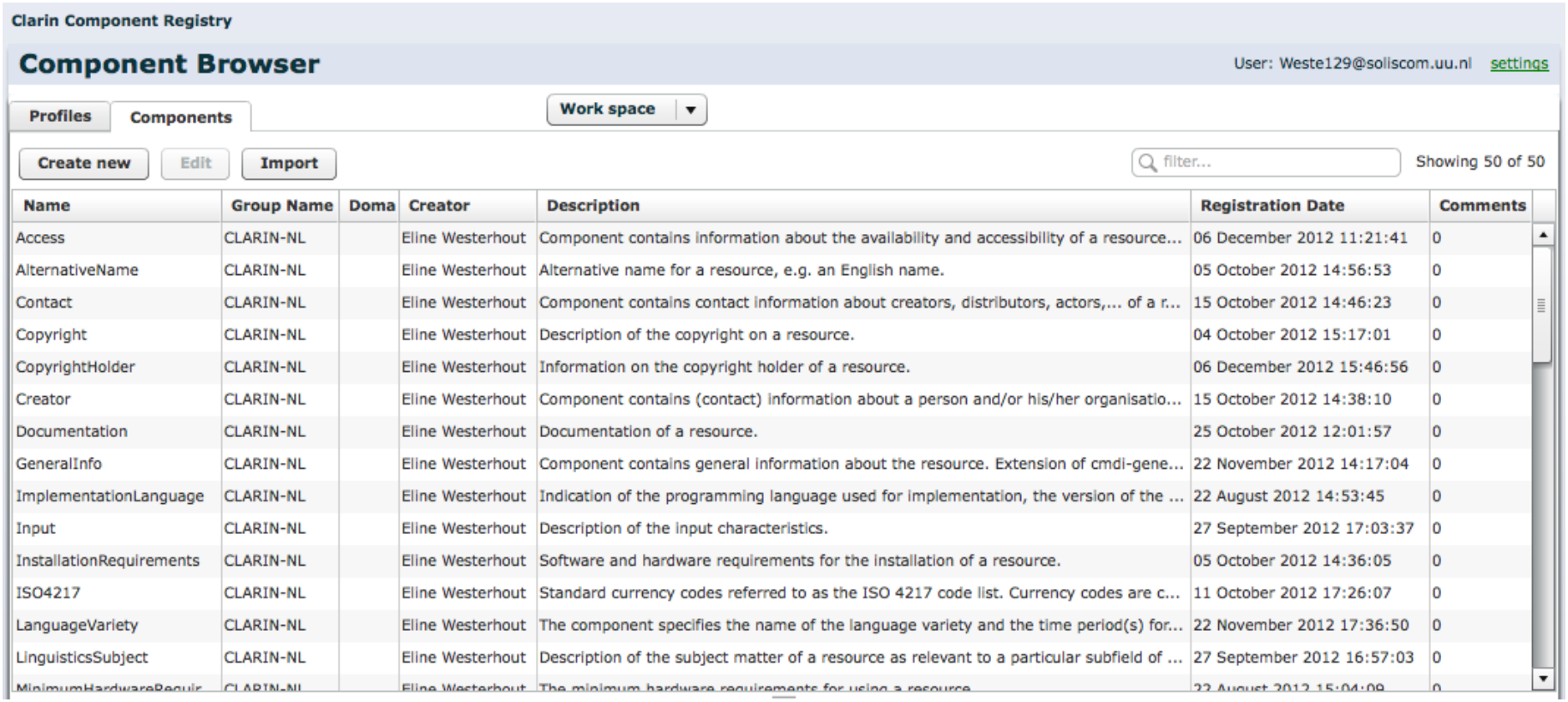
Implementation process



1. Identify concepts & structure

- input:
  - websites
  - metadata
  - manuals/reports
- structure:
  - GeneralInfo
  - SoftwareFunction
  - SoftwareImplementation
  - Access
  - Documentation
  - SoftwareDevelopment
  - TechnicalInfo
  - Service

2. Build components



Name	Group Name	Doma	Creator	Description	Registration Date	Comments
Access	CLARIN-NL		Eline Westerhout	Component contains information about the availability and accessibility of a resource...	06 December 2012 11:21:41	0
AlternativeName	CLARIN-NL		Eline Westerhout	Alternative name for a resource, e.g. an English name.	05 October 2012 14:56:53	0
Contact	CLARIN-NL		Eline Westerhout	Component contains contact information about creators, distributors, actors,... of a r...	15 October 2012 14:46:23	0
Copyright	CLARIN-NL		Eline Westerhout	Description of the copyright on a resource.	04 October 2012 15:17:01	0
CopyrightHolder	CLARIN-NL		Eline Westerhout	Information on the copyright holder of a resource.	06 December 2012 15:46:56	0
Creator	CLARIN-NL		Eline Westerhout	Component contains (contact) information about a person and/or his/her organisatio...	15 October 2012 14:38:10	0
Documentation	CLARIN-NL		Eline Westerhout	Documentation of a resource.	25 October 2012 12:01:57	0
GeneralInfo	CLARIN-NL		Eline Westerhout	Component contains general information about the resource. Extension of cmdi-gene...	22 November 2012 14:17:04	0
ImplementationLanguage	CLARIN-NL		Eline Westerhout	Indication of the programming language used for implementation, the version of the ...	22 August 2012 14:53:45	0
Input	CLARIN-NL		Eline Westerhout	Description of the input characteristics.	27 September 2012 17:03:37	0
InstallationRequirements	CLARIN-NL		Eline Westerhout	Software and hardware requirements for the installation of a resource.	05 October 2012 14:36:05	0
ISO4217	CLARIN-NL		Eline Westerhout	Standard currency codes referred to as the ISO 4217 code list. Currency codes are c...	11 October 2012 17:26:07	0
LanguageVariety	CLARIN-NL		Eline Westerhout	The component specifies the name of the language variety and the time period(s) for...	22 November 2012 17:36:50	0
LinguisticsSubject	CLARIN-NL		Eline Westerhout	Description of the subject matter of a resource as relevant to a particular subfield of ...	27 September 2012 16:57:03	0
MinimumHardwareRequir	CLARIN-NL		Eline Westerhout	The minimum hardware requirements for using a resource	22 August 2012 15:04:09	0

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.dfki.de/)
DisplayPriority: 1
Number of occurrences: 1 - unbounded

Element: **toolTask** abbreviationExpansion
ConceptLink: http://www.isocat.org/datcat/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

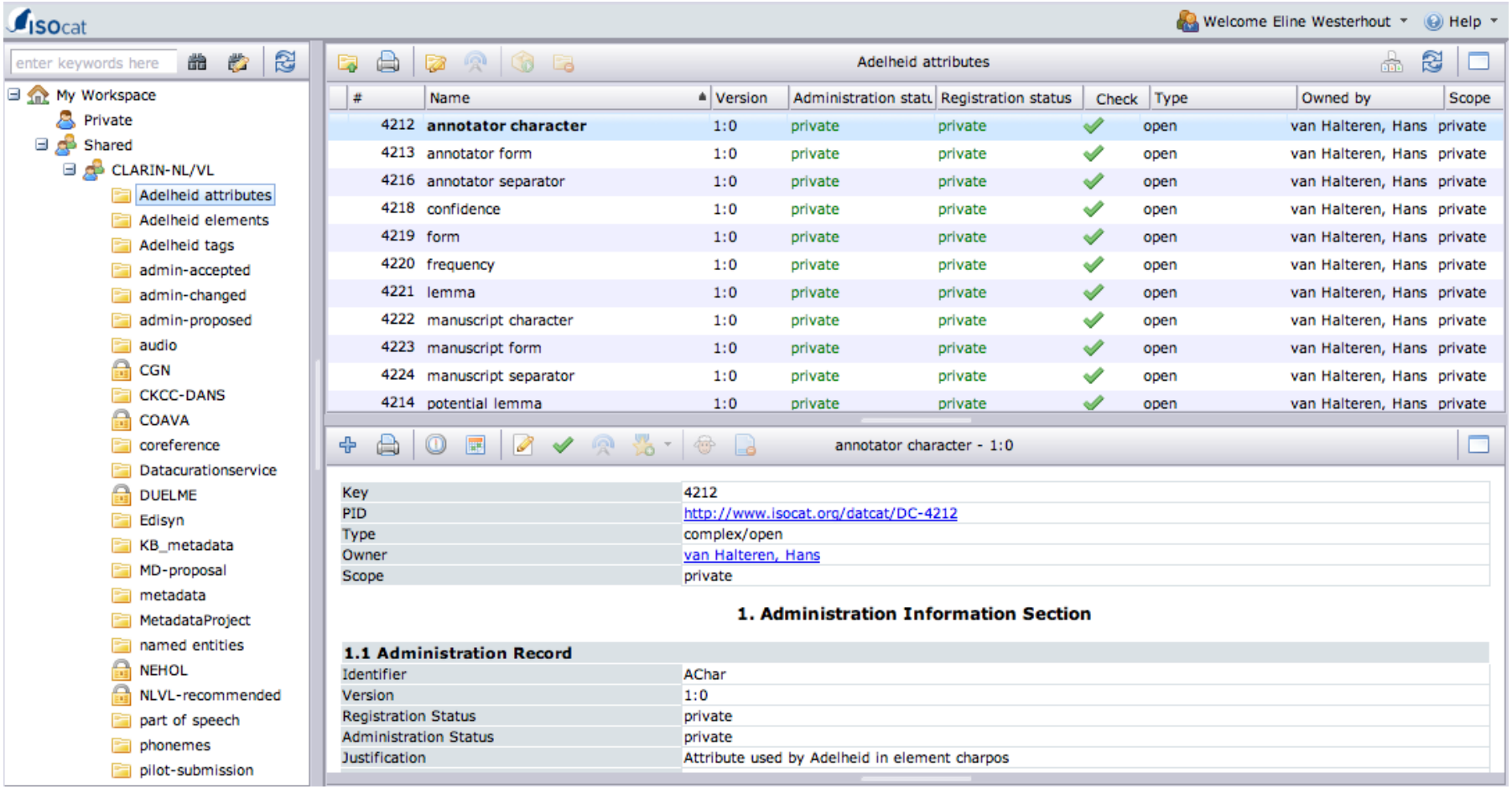
Component: **LinguisticsSubject**
Number of occurrences: 1 - unbounded

Component: **LanguageVariety**
Number of occurrences: 1 - unbounded

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

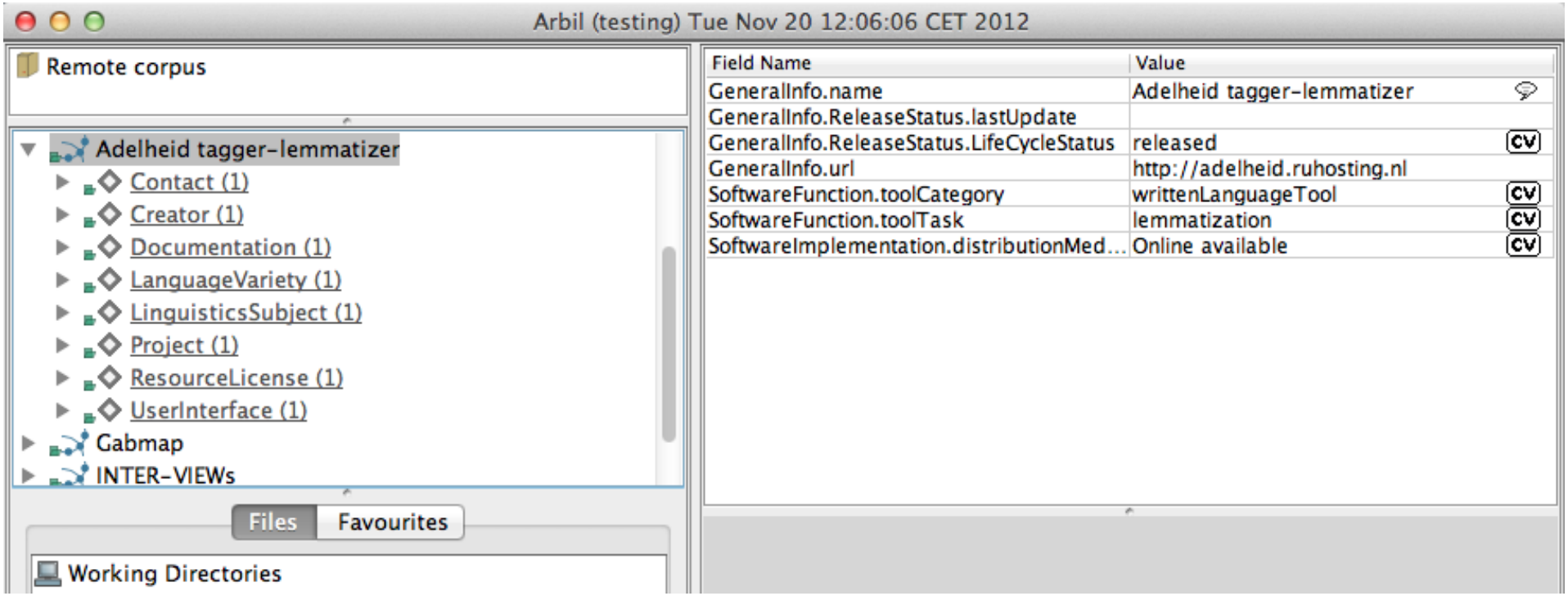
Component: **cmdi-description**
Number of occurrences: 0 - 1

3. Link to ISOcat



4. Build profile

5. Fill in metadata - I



5. Fill in metadata - II

```
<?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"
     xsi:schemaLocation="http://www.clarin.eu/cmd/ http://catalog.clarin.eu/ds/
           ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1342181139640/xsd">
   <Header>
      <MdCreator>elinewesterhout</MdCreator>
      <MdCreationDate>2013-01-11+01:00</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>
            <LanguageVariety>
               <century>14</century>
               <languageDependent>1</languageDependent>
               <ISO639>
                  <iso-639-3-code>nld</iso-639-3-code>
               </ISO639>
            </LanguageVariety>
         </SoftwareFunction>
         <SoftwareImplementation> ... </SoftwareImplementation>
         <Access> ... </Access>
         <ResourceDocumentation> ... </ResourceDocumentation>
         <SoftwareDevelopment> ... </SoftwareDevelopment>
      </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