Digital Objects & CLARIN use cases

Dieter Van Uytvanck, Willem Elbers
GEDE DO Workshop
26 September 2018
Brussels, Belgium
CLARIN in six bullets

• **CLARIN** is the Common Language Resources and Technology Infrastructure
• **ESFRI** ERIC status since 2012, Landmark since 2016
• that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
• to **digital language data**
  – in written, spoken, video or multimodal form
• and **advanced tools**
  – to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
• through a **single sign-on** online environment
CLARIN ERIC in members and centres

A consortium of:

20 members: AT, BG, CZ, DE, DK, DLU, EE, FI, GR, HR, HU, IT, LT, LV, NL, NO, PL, PT, SE, SI

2 observers: FR, UK;

>40 centres
Data architecture at a CLARIN centre

Repository at a CLARIN centre

Language Data

Metadata

Language Tools

Single text or recording
Corpus
Lexicon
Wordnet
Grammar

Web application
Web service
Web service pipeline
Stand-alone application

...
The CLARIN data architecture: Persistent Identifiers (PIDs)
Data discovery and processing

Harvested Metadata → Metadata Search → Data Analysis

Indexation → Selection

Virtual Language Observatory

- Language Data
- Metadata
- Language Tools
Language Resource Switchboard

Match based on language, format & task
Language Resource Switchboard

CLARIN & EOSC-hub
- Via metadata
- Virtual Language Observatory
- Virtual Collection Registry

EUDAT
- Via metadata
- B2DROP [cloud storage]

Parthenos
- Via metadata
- D4Science [Virtual Research Environment]

C2CAMP
- Via direct input of digital objects
  Matching based on mime type and language information stored in handle
CLARIN Digital Object requirements

<table>
<thead>
<tr>
<th>Persistent Identifier</th>
<th>Metadata Identifier</th>
<th>Mimetype</th>
<th>Language</th>
<th>License?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Data</td>
<td>Metadata</td>
<td>text/plain</td>
<td>ISO639-3: nld</td>
<td>GPL v3</td>
</tr>
</tbody>
</table>

Describes
Virtual Collections

Collection

Metadata

Metadata | Object
--- | ---

Metadata | Object
--- | ---

Metadata | Object
--- | ---

Referencing object via URL
Object metadata managed by collection owner

Referencing object via PID
Object metadata managed by collection owner

Referencing object via Digital Object Identifier
Object metadata collected via DO protocol

URL

PID

Digital Object
Digital Object Protocol

Uniform interface to object:
- DO Metadata
- Bitstream
- Descriptive metadata

Language Resource Switchboard
Virtual Collection Registry
Other services
Summary

• CLARIN use-cases
  - Language Resource Switchboard
  - Virtual Collection Registry
  - DO Protocol for repositories

• Important DO properties
  - Mime Type
  - Language
  - License
  - Reference to descriptive / user metadata