How can I find resources using CLARIN?

Dieter Van Uytvanck
CLARIN ERIC
dieter@clarin.eu

Using CLARIN for digital research, DH Lausanne
2014-07-07
Overview

- Resources?
- A closer look at the Virtual Language Observatory
  - Facet browser
  - Google Earth overlay
- Federated login to protected resources
Language resources?

- Broad term, covering:
  - Data collections
    - Corpora
    - Lexica
    - Grammars
    - Multimedia recordings
  - Software
  - Web applications / services
  - Digitized:
    - Books
    - Articles
    - Minutes of parliamentary debates
- ... 
- It’s like a jungle, sometimes ...
VLO: the idea

- Researcher: “where do I start?”
- Provide a single entry point giving access to all information
- Because of the large amount of data:
  - Drill-down paradigm (decrease search space gradually)
  - Multiple ways of exploring:
    - Full-text search
    - Facet browsing
    - Geographic overlay
  - Unified interface, links to the original context
VLO?

- Virtual Language Observatory
- http://www.clarin.eu/vlo/
- Several parts:
  - **Facet browser** (real search)
  - **Google Earth overlay** (geographic overview)
  - LRT inventory (ad-hoc, last resort metadata entry)
Facets?

- A **simple** way to narrow down the search space, step by step
- Values offered are **dynamic**: they change with every previous selection made
- Purpose: **quickly navigating** through a huge amount of metadata
- Gives an idea of what is “out there”
- Useful too for **metadata curation**
- **Not** the tool to answer research questions!
"PolDiLemma" Middle Polish Diachrone Lemmatised Corpus

The PolDiLemma corpus is a diachronic corpus made of political, religious, scientific and historical texts from different authors of the Middle Polish period (16th-18th century). Characteristic for this period is the slow development of a supra-regional standard language, a process of standardisation on the basis of t...

A Latin Grammar

A Valency Lexicon for Latin extracted from the Latin Dependency Treebank

This resource is a corpus-based valency lexicon for Latin verbs, automatically extracted from the Latin Dependency Treebank. It contains the full list of patterns of arguments occurring with each verbal form in the Latin Dependency Treebank, with syntactic information on the dependency role of the arguments, morphologi...

A collection of ten samples of Latin from various works of Francis Bacon

[Electronic resource]
## BREVIAVIRIAM AB URBE CONDITA [ELECTRONIC RESOURCE] / EUTROPIUS

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Breviariam ab urbe condita [Electronic resource] / Eutropius</td>
</tr>
<tr>
<td>Collection</td>
<td>Oxford Text Archive →</td>
</tr>
<tr>
<td>Language</td>
<td>Latin</td>
</tr>
<tr>
<td>Organisation</td>
<td>Oxford Text Archive, University of Oxford →</td>
</tr>
<tr>
<td>Resource type</td>
<td>Text →</td>
</tr>
</tbody>
</table>

### RESOURCES

- [0330.xml](#)
All metadata used is in CMDI format

Metadata sources
  - CMDI files harvested from CLARIN centres
  - CMDI’fied OLAC records (from CLARIN centres and others)
  - CMDI’fied LRT inventory records

You can get to resources directly from search results

It’s also possible to add your resources, see the FAQ at clarin.eu/vlo
Some examples

- [www.clarin.eu/vlo](http://www.clarin.eu/vlo)

- **Find some resources in the catalogue:**
  - Corpus Gysseling
  - Telephone conversation recordings in Nepal

- **Find some resources in the Endangered Languages archive which are:**
  - (Spoken) discourse with at least two consultants in Asia
  - Or (spoken) discourse with at least two consultants in a Face to Face conversation
Still to come…

- A faceted browser is as good as its (meta)data:
  - so curation steps are needed
  - more metadata (e.g. from national libraries)
    http://catalog-clarin.esc.rzg.mpg.de/vlo/
- Some more facets under investigation: e.g. time
- Interface improvements
  - E.g. better support for touch screens
Inherent limitations: Simple search
- no OR combinations possible (not yet…)
- no sophisticated search operations
- no visualisation of complex relations (e.g. hierarchical collections)

Scope: the VLO is not Google
- But: more precise selection, integration with CLARIN infrastructure

Heterogeneous metadata sources
- Metadata language
- Different ways of expressing links to data:
  - direct download
  - via a landing page
  - maybe only a request form > access rights
Federated login

- General tendency towards open access, however there are still factors as:
  - Scientific competition
  - IPR
  - Privacy and ethical issues
- If password-protection is needed at least it should be made simple (single sign-on)
- Access a list via
  - http://clarin.eu/node/3962
  - CLARIN site > Services > Easy access to protected resources
Google Earth overlay

- Collection of placemarks related to language resources
- Contains:
  - Languages, with links to
    - Wikipedia
    - Linguist List Map
    - Ethnologue
    - WALs typological database
    - The VLO facet browser
  - Selected examples of field research sites
Questions?

- … ask them now
- … or send a mail to vlo@clarin.eu

More information:
- www.clarin.eu/vlo
- www.clarin.eu/cmdi