CLARIN-IT Resources in CLARIN ERIC
A Bird’s-Eye View

Dario Del Fante
ILC-CNR - Italy
dario.delfante@ilc.cnr.it

Francesca Frontini
ILC-CNR - Italy
francesca.frontini@ilc.cnr.it

Monica Monachini
ILC-CNR - Italy
monica.monachini@ilc.cnr.it

Valeria Quochi
ILC-CNR - Italy
valeria.quochi@ilc.cnr.it

CLARIN Conference 2021
CLARIN - The technical infrastructure

Harvested Metadata

Metadata Search

Virtual Language Observatory (VLO)

Data Analysis

Language Resource Switchboard (LRS)

clarin.eu/fair

vlo.clarin.eu

switchboard.clarin.eu

Corpora
- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools
- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

Spoken corpora in the CLARIN infrastructure
Corpora with transcriptions and audio recordings

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th>Description</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic Speech Corpus</td>
<td>Arabic</td>
<td>The corpus is available for download from a dedicated website.</td>
<td>Download</td>
</tr>
<tr>
<td>SIMPLE: English spoken corpora with multi-transcription</td>
<td>English</td>
<td>For a related publication, see reference (2019).</td>
<td>Download</td>
</tr>
<tr>
<td>Czech</td>
<td>Czech</td>
<td>This corpus contains traditional dictation material, mostly unprepared monologues. For a related publication, see reference (2019).</td>
<td>Download</td>
</tr>
</tbody>
</table>

https://www.clarin.eu/resource-families
Research aims

National consortia should monitor regularly these four “points of access”

A more qualitative assessment to ensure that any researcher/end-user can easily find the resources she/he needs and use them as intended.

A methodology to check and analyze the presence of the LRs available in the CLARIN-IT consortium:

• Assessing the Italian consortium presence
• Devising a reproducible qualitative methodology from the user perspective.
• Evaluating the visibility, reliability and searchability of CLARIN-IT LRs in the VLO.
CLARIN-IT - Technology infrastructure

CLARIN
Common Language Resources and Technology Infrastructure

LINDAT
CLARIN-CE

GARR cloud

Data registration

EDUDAT

SSO

idem garr aai

eduGAIN

SWE-CLARIN

ERCC
The Methodology

Select in the National Project tab – focus on LRs of interest

Check which LRs are shown and how are presented in the VLO, filtering for:
- Languages
- Organisation Collections
- Resource type

Check the presence of duplicates

Check the status of activation for a sample of links to the original place

Register all the inconsistencies in terms of accessibility and availability
CLARIN-IT – A Bird’s eye view

<table>
<thead>
<tr>
<th>CLARIN-IT - a birds eye view</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Number of LR</strong></td>
<td>Latin 366</td>
</tr>
<tr>
<td><strong>Monolingual</strong></td>
<td>English 40</td>
</tr>
<tr>
<td><strong>Multilingual resources</strong></td>
<td>Arabic 32</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>Ancient Greek (to 1453) 6</td>
</tr>
<tr>
<td><strong>Languages</strong></td>
<td>Ancient Greek 8</td>
</tr>
<tr>
<td><strong>Organisations</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Collections</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Resource type</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Data providers</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: CLARIN-IT on VLO

<table>
<thead>
<tr>
<th>Organisations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archivio della Latinità Italiana del Medioevo (ALIM) 354</td>
</tr>
<tr>
<td>Istituto di Linguistica Computazionale - CNR 39</td>
</tr>
<tr>
<td>Institute for Applied Linguistic Research - EURAC 9</td>
</tr>
<tr>
<td>Università di Salerno 8</td>
</tr>
</tbody>
</table>

Table 3: Organisations in CLARIN-IT
## CLARIN-IT – A Bird’s eye view

<table>
<thead>
<tr>
<th>Collections</th>
<th>Count</th>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALIM Literary Sources</td>
<td>344</td>
<td>ILC4CLARIN : OPEN Data and Tools</td>
<td>7</td>
</tr>
<tr>
<td>ILC4CLARIN</td>
<td>54</td>
<td>ERCC Learner Corpora</td>
<td>8</td>
</tr>
<tr>
<td>Alim Documentary Sources</td>
<td>11</td>
<td>ERCC Web Corpora</td>
<td>4</td>
</tr>
<tr>
<td>CIRCSE</td>
<td>8</td>
<td>ERCC</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 4: Collections in CLARIN-IT**

<table>
<thead>
<tr>
<th>ILC4CLARIN</th>
<th>Eurac Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>368</td>
</tr>
<tr>
<td>Lexical Resource</td>
<td>43</td>
</tr>
<tr>
<td>Software, webservice</td>
<td>12</td>
</tr>
<tr>
<td>Webservice</td>
<td>2</td>
</tr>
<tr>
<td>Text</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 5: Resource type for each Data provider**
Concluding remarks

**Availability** - cases of unspecified availability might lead to amendments of the records

**Granularity** – consider nested vs non-nested options when preparing a resource

**Duplicate Namings** – pay attention to differentiate those resources corresponding to the same work with different editors.