The FAIR Principles in CLARIN

Franciska de Jong and Dieter Van Uytvanck

CLARIN ERIC

Joint PARTHENOS – CLARIN - DARIAH Workshop on Applying the FAIR Principles

Utrecht

6 July 2017
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:

• 19 members: AT, BG, CZ, DE, DK, DLU, EE, FI, GR, HU, IT, LT, LV, NL, NO, PL, PT, SE, SI
• 2 observers: FR, UK;
• >40 centres
CLARIN in data types

- Parliamentary records
- Literary texts
- Social Media data
- Historical letters
- Oral History data
- Disciplinary libraries
- Institutional archival data
- Broadcast archives
- Newspaper archives
- …
CLARIN and data science

- Analytics for text and speech data as a pillar for data science
- Contribution to the development of new methodological frameworks for the integrated processing of multiple datatypes and multidisciplinary research agendas.
- Europe’s multilinguality as a basis for comparative research of societal and cultural phenomena, and in particular those that are reflected in language use; some examples:
  - Migration patterns
  - Intellectual history
  - Language variation across period and region
  - Dynamics in mental health conditions
  - Parliamentary discourse
- Text and speech as social and cultural data
- From tools for the study of lexical units to big data analysis tools
CLARIN centres

- A **distributed architecture**: (http-accessible) files, web applications and web services spread all over Europe
- Nodes in the network: **centres**
- Currently 40 registered centres, of which 20 certified B-centres
The CLARIN data architecture: repositories

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
  ...

describes
The CLARIN data architecture: harvesting

Diagram showing the relationships between Language Data, Metadata, and Language Tools, with Harvested Metadata at the center.
The CLARIN data architecture: processing
The CLARIN data architecture: content search

(Federated) Content Search

1. Enter query
2. Perform local search
3. Retrieve results
4. Show aggregated results
The CLARIN data architecture: workflows

Web Service Pipelines
(1) select input data
(2) construct pipeline
(3) execute
(4) use/analyse output data
The CLARIN data architecture: Persistent Identifiers (PIDs)
The CLARIN data architecture: Component Metadata (CMDI)

Metadata **Instance**

- Fixed Header
- Flexible Body

is based on one

Metadata **Profile**

- Metadata Components
FAIR

- [https://www.force11.org/fairprinciples](https://www.force11.org/fairprinciples)
  Data as increasingly FAIR Digital Objects
Findable

(meta)data are assigned a globally unique and eternally persistent identifier. OK

handles required for metadata handles required for data

F2 data are described with rich metadata. OK required what about granularity?

F3 (meta)data are registered or indexed in a searchable resource. OK

F4 metadata specify the data identifier. OK what about landing pages without direct data access?
## Accessible

### A1
(meta)data are retrievable by their identifier using a standardized communications protocol.

<table>
<thead>
<tr>
<th>A1.1</th>
<th>the protocol is open, free, and universally implementable.</th>
<th>OK handle + HTTP protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>the protocol allows for an authentication and authorization procedure, where necessary.</td>
<td>OK SAML for federated login</td>
</tr>
</tbody>
</table>

### A2
metadata are accessible, even when the data are no longer available.

OK (but: awareness is important!)
## Interoperable

| I1 | (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. | OK | metadata: CMDI data: recommended formats |
| I2 | (meta)data use vocabularies that follow FAIR principles. | OK | metadata > metadata |
| I3 | (meta)data include qualified references to other (meta)data. | OK | metadata > metadata > data |
### Re-usable

<table>
<thead>
<tr>
<th>R1</th>
<th>(meta)data have a plurality of accurate and relevant attributes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1.1</td>
<td>(meta)data are released with a clear and accessible data usage license.</td>
</tr>
<tr>
<td>R1.2</td>
<td>(meta)data are associated with their provenance.</td>
</tr>
<tr>
<td>R1.3</td>
<td>(meta)data meet domain-relevant community standards.</td>
</tr>
</tbody>
</table>
Points of attention

• Machine accessibility requires attention:
  – direct URLs/PIDs need to be included in metadata
  – data type needs to be included (e.g. mimetype)
  – landing page is not sufficient

• Fancy (but idiosyncratic) additions can make a repository less FAIR, e.g.:
  – special viewers (if they hide or block direct access to the data)
  – dialogs, e.g. to agree on terms of use
  – non-standardized login workflows
Conclusions

• Being FAIR-compliant costs considerable effort (more than one might foresee at first sight!), however:
  – It helps a lot when being FAIR is considered while designing the data architecture
  – For us, the results (user satisfaction, multiplier effects) were worth the work

• With the recent visibility of FAIR, it is good to keep in mind that they are guiding principles: not a religion, nor an empty buzzword.
Thank you for your attention!

• Questions?

• For more details:
  – CLARIN centres overview: https://www.clarin.eu/content/clarin-centres
  – CLARIN centre requirements: http://hdl.handle.net/11372/DOC-77
  – Force11 on FAIR Principles: https://www.force11.org/fairprinciples