Data and Computing infrastructures for Global Linguistic Resources: The Scientist Perspective

Erhard Hinrichs
CLARIN ERIC Co-Director
erhard.hinrichs@uni-tuebingen.de
2015-09-22

Workshop Day, RDA 6th Plenary, Paris
CLARIN?

- Common Language Resources and Technology Infrastructure
- European (ESFRI) Research Infrastructure – ERIC since February 2012
- aims at providing easy and sustainable access for scholars in the humanities and social sciences
  - to digital language data (in written, spoken, video or multimodal form)
  - to advanced tools to discover, explore, exploit, annotate, analyse or combine them
CLARIN Resources and Services

- Access to language resources (including federated login when needed)
- Access to language resource processing applications/services
- Depositing services
- Metadata catalogue: Virtual Language Observatory
- Federated Content Search
- “Glue components” like the Virtual Collection Registry (http://clarin.eu/vcr)
- Consulting services

Visit: http://clarin.eu/services
CLARIN in Europe

- **Members:** (recently joined)
  - Austria
  - Bulgaria
  - Czech Republic
  - Denmark
  - Dutch Language Union (Belgium, NL)
  - Estonia
  - Germany
  - Greece
  - Lithuania
  - Netherlands
  - Poland
  - Portugal
  - Slovenia
  - Sweden
  - Norway

- **Candidates (Observers):**
  - Finland
  - United Kingdom
The 28 CLARIN centres
Rome and Paris Papers

Both documents are very valuable for CLARIN:

- **Rome paper**: *Future of research data and computing infrastructures*
  - call for actions identifies three reflection points highly relevant for CLARIN

- **Paris paper**: *Data Management Trends, Principles, and Components*
  - good reference for CLARIN members and centers
  - treats research data not as static objects, but takes the production process of data via workflows into account
Metadata descriptions and PIDs for digital objects as input/output in self-documenting workflows

see Paris paper; Figure 7, p. 6
Reproducible Science in CLARIN

The Potsdam/Leipzig **Mind Research Repository (MRR)**

- User communities: cognitive psychology and cognitive neuroscience

- Provides access to publications along with data and scripts for analyses and figures reported in them

- [http://openscience.uni-leipzig.de/index.php/mr2](http://openscience.uni-leipzig.de/index.php/mr2)
Workflow Engines in CLARIN

WebLicht
- Linguistic annotation of user input:
  - Tokens
  - POS Tags
  - Lemmas
  - Semantics
  - Constituent Parse trees
  - Dependency Parse trees
  - Named Entities
- (Statistical) analysis, visualization, ISOcat integration
- Converters to linguistic standard formats

Flexible user-defined workflows
Typical CLARIN User Scenario

- A historian, linguist, or political scientist
  - makes use data sets and web services hosted in one or more CLARIN Centers
  - generates one or more derived data set (e.g. as a virtual collections) and published his analysis of the derived data set

- To make the published analysis reproducible, reference is needed to
  - the version(s) of the data sets used
  - the version of the tool(s) and tool parameters used for the analysis
Practical Consequences

- Executables of the tools need to be stored by the scientist or the tool provider.
- Global PIDs to the data sets involved are often not sufficient.
- Reference data sets with new PIDs may need to be created.
Rome Reflection Points

- How the world of science and innovation is changing with data and computing infrastructures?

  Establishment of best practices for researchers, data and tool providers, funders, and publishers

- What are the gaps and needs to promote better and more social-aware science?

  Developing participatory institutional and funding models for user communities, universities, research institutes, data/tool providers, and data centers.
Thank you for your attention!

For more details:

http://clarin.eu > Services / Centres / Technology
CLARIN Centres

- A **distributed architecture**: (http-accessible) files, web applications and web services spread all over Europe
- Nodes in the network: **centres** ([http://clarin.eu/centres](http://clarin.eu/centres))
- Some of them password-protected (licenses, privacy, …)
- User base: also spread over Europe (and rest of the world)
DFT in CLARIN’s architecture

- **digital object**
- **bitstream**

Diagram:
- Language Data
- Metadata
- Language Tools

Describes:
DFT in CLARIN’s architecture

- Language Data
- Metadata
- Language Tools

- single text or recording
  - corpus
  - lexicon
  - wordnet
  - grammar
  - ...

- describes

- web application
  - web service
  - web service pipeline
  - stand-alone application
  - ...

...
DFT: Access Methods

1.2.2 Access Methods
We can basically identify 2 methods to access a bitstream that encodes the information we ...
DFT in CLARIN’s architecture: Virtual Collections

my collection

digital collections

Registry
DFT in CLARIN’s architecture: Virtual Collections

- VC registry available
- http://clarin.eu/content/virtual-collections

**Absolute spatial deixis and proto-toponyms in Kata Kolok**

**General**
- **Name:** Absolute spatial deixis and proto-toponyms in Kata Kolok
- **Type:** extensional
- **Creation Date:** 2014-09-26
- **Purpose:** research
- **Reproducibility:** intended
- **Persistent identifier:** hdl:11372/VC-1001
- **Keywords:** • sign language
  • Kata Kolok

**Creators**
- **Person:** Connie de Vos
- **Organisation:** Max Planck Institute for Psycholinguistics
- **Website:** http://www.mpi.nl/people/vos-connie-de
- **Role:** Researcher

**Resources**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Type</th>
</tr>
</thead>
</table>
| Journal Article (fulltext)
This paper presents an overview of spatial deictic structures in Kata Kolok, a sign language which is indigenous to a Balinese village community. | Resource |
| Footnote 3 - video
Absolute versus absolute transpositional pointing signs | Resource |
| Footnote 4 - video
COME-HERE-FROM-A and GO-FROM-HERE-TO-B | Resource |
Feedback as early adopter

- Very handy to have a lingua franca when discussing research infrastructure architectures
- It was good to be involved as adopting community from the start of the work
- Would be good to have a single and concise summarizing document / webpage (with a PID :)
DFT in CLARIN’s architecture

**metadata**

**persistent ID**