CLARIN 2021
Repositories and National CLARIN Centres

Chair: Neeme Kahusk
Day 2
Tuesday 28 September
12:30 - 12:55
ARCHE Suite:
A Flexible Approach to Repository Metadata Management

Mateusz Żółtak, Martina Trognitz and Matej Durco
mateusz.zoltak@oeaw.ac.at, martina.trognitz@oeaw.ac.at, matej.durco@oeaw.ac.at
ACDH-CH
Why a new solution?

- Native RDF support
- Extensible (without Java)
- Support for multiple identifiers
- Reliable transactions
- Reliable backups
- Fedora 6
- Fedora 4/5
- Invenio
- DSpace
- Dataverse
- Search API
Why do we need RDF?

• To migrate easily from Fedora 4
• Because we like LOD and Semantic Web
• RDF provides a nice way (LOD and SW compliant) of dealing with named entities
How do we deal with the RDF? 1/2

- Repository is just an RDF graph (repository resource <=> node in RDF graph)
  - Each node can optionally store a binary content

- Automatic "same as" detection on metadata ingestion to avoid duplicated nodes
How do we deal with the RDF? 2/2

Easy way to fetch all metadata of interest at once, e.g.

- metadata of a resource and all its RDF neighbors
- metadata of a resource and all its RDF "relatives"
CLARIN - ready

Support for

- FCS 2 basic search
- Switchboard
- VCR

Rich and fast OAI-PMH templating engine

- Allows on-the-fly XML metadata generation from the repository's RDF metadata
- Examples ([basic CMDI](https://example.com), [ARIADNE](https://example.com), [DHA](https://example.com), [Kulturpool](https://example.com))
Thank you for your attention!
A data repository for the management of dynamic linguistic datasets

Thomas Gaillat¹, Leonardo Contreras Roa², Juvénal Attoumbré¹

¹Université de Rennes, ²Université de Picardie Jules Verne

CLARIN Annual Conference 2021
1. **Introduction**

Most available language corpora are static and risk obsolescence.
Proposal: using a dynamic database to provide a seamless workflow from data ingestion to data querying and data set creation
2. Current Practices

- Use of single persistent URLs for entire datasets
  - Full downloads from repositories - Ex: Ortolang (Huma-Num)

- A move towards interoperability
  - NLP pipelines
  - Interoperable tools and datasets - Ex: European Language Grid
3. Use case: A data repository for a learner corpus

Corpus CIL

- English and French L2
- 115 speakers but 84 recordings in DB
- Spoken and written modules

Repository:
Data are organised in collections of data items
Each data item is assigned a persistent URL
Data items are grouped in collections
Each collection is assigned a persistent URL as well

<table>
<thead>
<tr>
<th>Corpus Inter Langue</th>
<th>Persistent URL</th>
<th>Persistent URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data item</td>
<td>Persistent URL</td>
<td></td>
</tr>
<tr>
<td>Learner data files (.wav,.pdf,.txt,.eaf) + learner metadata file (.csv)</td>
<td>Persistent URL</td>
<td>Dublin Core Metadata</td>
</tr>
<tr>
<td>Dublin Core Metadata</td>
<td>Persistent URL</td>
<td></td>
</tr>
</tbody>
</table>

Structure of CIL in Nakala:
One learner = one data item
One data item <- Dublin Core metadata + data files (incl. learner metadata file)
3. Use case: A data repository for a learner corpus

Generating linguistic datasets

- Data ingestion
- APIs
- Nakala

Data querying
- POS annotation
  - *UDPipe*
- Phonetic alignment
  - *WebMAUS*
- etc.

Enriched Linguistic Datasets

.csv
4. Perspectives

- development of customised scripts (R) for data annotation
  - semantic, morphosyntactic, phonological and phonetic
- using annotated data for didactic purposes
  - predicting learner proficiency level, determining weaknesses and strengths
- need for global state versioning of corpus
  - more precise version bookkeeping
  - longitudinal comparisons within one same corpus
Thank you for your attention!
References


Jan Wijffels (2020). *Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit.* R package version 0.8.5.


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>Analytic Speech Corpus</td>
<td>Analytic</td>
<td>This corpus is available for download from a dedicated website.</td>
<td>Download</td>
</tr>
<tr>
<td>SHARC</td>
<td>Czech</td>
<td>This corpus contains traditional data (e.g., written) and subcorpora (e.g., spoken) and is available for research.</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
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>Total Number of LR</td>
<td>Latin 366</td>
</tr>
<tr>
<td>Monolingual</td>
<td>Italian 30</td>
</tr>
<tr>
<td>Multilingual resources</td>
<td>English 40</td>
</tr>
<tr>
<td>Format</td>
<td>German 8</td>
</tr>
<tr>
<td>Languages</td>
<td>Arabic 32</td>
</tr>
<tr>
<td>Organisations</td>
<td>Czech 2</td>
</tr>
<tr>
<td>Collections</td>
<td>Ancient Greek (to 1453) 6</td>
</tr>
<tr>
<td>Resource type</td>
<td>Breton 1</td>
</tr>
<tr>
<td>Data providers</td>
<td>Ancient Greek 8</td>
</tr>
<tr>
<td></td>
<td>Basque 1</td>
</tr>
</tbody>
</table>

Table 2: Languages in CLARIN-IT

<table>
<thead>
<tr>
<th>Organisations</th>
<th>Total</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Archivio della Latinità Italiana del Medioevo (ALIM)</td>
<td>354</td>
<td>CIRCSE - Università Cattolica Sacro Cuore</td>
<td>8</td>
</tr>
<tr>
<td>Istituto di Linguistica Computazionale - CNR</td>
<td>39</td>
<td>Ghent Universities</td>
<td>2</td>
</tr>
<tr>
<td>Institute for Applied Linguistic Research - EURAC</td>
<td>9</td>
<td>Università di Parma</td>
<td>2</td>
</tr>
<tr>
<td>Università di Salerno</td>
<td>8</td>
<td>Basque</td>
<td>1</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>Provider</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.
Thank you for your attention!
Opening Language Resource Infra-structures to Non-Research Partners

Practicalities and Challenges

Egon W. Stemle | Verena Lyding | Alexander König
{egon.stemle, verena.lyding}@eurac.edu | alex@clarin.eu
Introduction | Domain Specific Infrastructure
Local Infrastructure | Aims

• Instead of targeting a specific domain (e.g., lexicography) target *local* stakeholders (use language resources in different ways)

• Connect *various types* of language actors on local level (exploit synergies between their activities and our goals*)

• Focus on non-research partners (usually not familiar with large infrastructure efforts)

* Our *overall* goal is to bring text producers together to gain access to authentic texts for corpus linguistic work
Local Infrastructure | Challenges

• *We* wanted (needed) someone to team up with us
• Invested time (= money) needs to produce revenue
• 3 out of 4 newspapers were not interested even though small monetary reward was offered

• Challenging to envision use cases that could be explored
  - knowledge of NLP technology is vague and sometimes overestimated
  - fear of disrupting established workflows and wary of *unnecessary* changes
  - abstract added value, like an evaluation metric, a promise of an improved experience, or a functioning prototype with different data
Local Infrastructure | salto.bz Use-Case

Cooperation with the local online newspaper portal salto.bz

• Improve search service for the text archive
• (Full-text + manually created metadata search)
• Automatic keyword suggestion (+ manual refinement)

• Implement technical part
  (mostly our side + directly involved people)
• Devise, implement, impose organisational part
  (mostly their side + indirectly involved people)
Wie erstelle ich Salto-Artikel? Wie füge ich Bilder ein?

Get Tags

Autotags

- @Eurac Research
- @Eurac Ricerca
- @Salto
- @Capriola
- @Schlagwortpaaren
- @Copple di parole chiave
- @Suchfunktion
- @Textdaten
- @Testo
- @Use Case
- @Caso d’uso

Autotags checked manually

Speichern Veröffentlichung zurücknehmen Änderungen ansehen
• Technical improvements – although important – are not an indicator of success

• Organisational integration and wide acceptance (also by those less involved) is a better indicator of (possible) long-term success

• Continuation of support requests after the end of the project

• More use-cases and (success) stories needed

• Innovative ideas for mid- and long-term support (financing) needed
Gesucht, gefunden


Von Editorial Salto 01.04.2021


Questions and Answers
CLARIN 2021
Research Data Management, Metadata and Curation (Part 2)

Chair: Jurgita Vaičenonienė

Day 2
Tuesday 28 September
12:55 - 13:15
Bagman - A Tool that Supports Researchers Archiving Their Data

Claus Zinn
Motivation

• RDM essential for good scientific practise
  - theory builds on data
  - theory verification also means data inspection

• RDM often feels like household chores
  - inventory of all data
  - cleaning-up data, structuring data
  - documentation of procedures and scripts
  - metadata provision
  - bundling and archiving

• Tool required to support RDM
Background

1. Data Packaging
   • thinking of common / bad practise…
   • Docuteam Packer (software), BagIt (specification)

2. Metadata Assignment
   • CMDI as metadata infrastructure framework
   • Complexity makes it hard to use to its full potential

3. Archiving Procedure
   • where should my data best go, e.g., to maximise visibility
Bagman

- Web-based software
- Supports Researchers
- Supports Archive Managers

Bagman helps you package your research data and transfer it safely to your archive manager.

Your data will be shared with the archive manager only. For more details, see the FAQ.
Bagman

- Metadata collection via forms
- Different categories
- Predefined profiles
- Fixed set of licences
Bagman – Demonstration

• [https://weblicht.sfs.uni-tuebingen.de/bagman/](https://weblicht.sfs.uni-tuebingen.de/bagman/)

• fully functional prototype version
  - only TALAR archive is connected
  - other archives, other profiles to be included
    • logins for archive managers
    • CMDI generation (and additional forms) for new archives and their requirements

• initial feedback encouraging
  - e.g., now browser session can be saved and later restored to continue with RDM
Bagman – Discussion

• design walks fine line between researchers and archive managers
  - for one side, the more metadata, the better
  - for the other side, get it done quickly…
  - some data mandatory, some other data can be added voluntarily

• no direct editing of CMDI
  - entire CMDI is generated from form-based input
  - and from the data being archived

• no ideological warfare
  - e.g., granularity of data being archived, or data formats being accepted
  - Bagman as the middle man, also as a means to establish communication
The TEI-based ISO Standard ‘Transcription of Spoken Language’ as an Exchange Format within CLARIN and Beyond

Hanna Hedeland and Thomas Schmidt
Leibniz-Institut für Deutsche Sprache - Universität Basel
Transcription of spoken language

No single stream of tokens, rather partly overlapping streams of tokens and non-token elements like pauses, non-phonological and non-verbal behaviour

Various theory-dependent transcription systems define tokens, non-tokens, linguistic units, annotations and general transcript layout

We can and should standardize non theory-dependent features (e.g. participants, time information, common information types) of transcription systems

GAT conventions (Selting et al. 1998):

MJ: I ((cough)) see a door. I (0.3) want to paint it (black/blue).

HIAT conventions (Rehbein et al. 2004):

MJ[v] I ((cough)) see a door. I ((0,3s)) want to paint it (black). MJ[k] (blue)
Enhanced Interoperability through the ISO/TEI Standard

There is already syntactic interoperability between transcription/annotation tool formats, the ISO/TEI Standard additionally offers basic **semantic interoperability** for transcription data.

GAT conventions
(Selting et al. 1998)

HIAT conventions
(Rehbein et al. 2004)
Some Use Cases in CLARIN and Beyond
TEI in CLARIN - FAIR enough?

Common challenges

• Linking the contents of (open) digital language resources
• Documenting (linguistic) annotation layers, schemas and categories
• Encoding and linking contextual information on recording sessions and participants or authors and background information
• Modeling TEI documents and TEI corpora as FAIR Digital Objects?
Thanks for your attention!

https://exmaralda.org/
http://zumult.ids-mannheim.de/ProtoZumult/
http://ltdemos.informatik.uni-hamburg.de/webanno-mm/
An Idea for Citation Tracking and Versioning for Linguistic Examples

Tobias Weber
Ludwig-Maximilians-Universität München
weber.tobias@campus.lmu.de
Goals

• Data citation tracking including publications from the past.
• Generation of graphs illustrating versions.

Note: This has not been implemented yet

• Building on prior work:
  - bit.do/WeberLDK
  - bit.do/WeberSKG2020
Why CLARIN?

• CLARIN has necessary infrastructures:
  - Computational power
  - Virtual collections; relational data and metadata formats
  - Front-end of Switchboard and VLO

• CLARIN has data:
  - Modern language data
  - Historical data sets prior to DOIs, URNs etc.

• CLARIN has political weight:
  - Negotiating access to publications
How?

• Parsing publications with string matching:
  - Versions of linguistic examples
  - Metadata

• Creation of separate entries in the collection:
  - Keep original form and metadata
  - Similarity analysis to establish relationships
  - (tentative) versions → graph

• Display to the public:
  - Different analyses of an example
  - Impact of research
Thank you for your attention

Get in touch

weber.tobias@campus.lmu.de
Questions and Answers

CLARIN ANNUAL CONFERENCE
Five-minutes paper presentations - chairs: N. Kahusk, J. Valčienienė, K. Lindén
28/09/2021 - 12:30

1. REPOSITORIES & NATIONAL CLARIN CENTRES
   - Repositories
   - National Clarin Centres
   - NAKALA
   - ARCHE SUITE
   - CLARIN-IT

2. RESEARCH, DATA MANAGEMENT, METADATA & CURATION - II
   - Bagman
   - Linguistics Citation Tracking

3. LEGAL ISSUES RELATED TO THE USE OF LRS IN RESEARCH - II
   - NLP
   - Twitter & Research
   - Copyright in Ethnomusicological Datasets

4. NLP FOR NON-RESEARCHERS

INSTAGRAM.COM/NONLINEARE
CLARIN 2021
Legal Issues Related to the Use of LRs in Research (Part 2)

Chair: Krister Lindén

Day 2
Tuesday 28 September
13:15 - 13:30
Legal Issues Related to the use of Twitter Data in Language Research

Pawel Kamocki, Vanessa Hannesschläger, Esther Hoorn, Aleksei Kelli, Marc Kupietz, Krister Lindén, Andrius Puksas
Twitter: an important data source

Twitter is publicly available, huge and still growing…

…and yet, few tweet corpora are widely known.

Is it because of the many legal gray zones?
Tweets are copyright-protectible

Maximum tweet length: 280 characters = 50-60 words, enough to meet the originality standard.

Many (most?) tweets are not original, but it is quite impossible sort them out automatically.

Copying tweets in bulk is a copyright-relevant act (reproduction is a restricted act).
Tweets contain personal data

Broad definition of personal data in the GDPR (‘any information related to an identifiable person’).

Twitter offers sophisticated privacy parameters, so public tweeting can possibly be interpreted as consent for re-use.

PROBLEM: It is difficult to delete a tweet from a corpus after the user deletes it from Twitter (withdraws consent)…
Twitter prohibits and enables mining

Twitter Terms of Service prohibit scraping tweets
- sanction: account suspension…?

BUT
Twitter provides an API for mining tweets, with a special track for researchers.

PROBLEMS:
1) in order to be granted access, one has to fill in a detailed questionnaire;
2) access can be terminated at any time for any reason (including no reason at all).
Alternative: the TDM exception?

The latest copyright directive contains an exception for TDM for research purposes:
- deadline for transposition: June 2021;
- overrides contracts (e.g. the prohibition in Twitter ToS).

PROBLEMS:
1) does not allow sharing (or only very limited);
2) not clear if it allows to override technological measures (protected by law).

Does the new exception really change the game between researchers and content providers like Twitter?
Thank you for your attention,
join us for the discussion!

kamocki@ids-mannheim.de (and co-authors)
Ethnomusicological Archives and Copyright Issues: An Italian Case Study

Prospero Marra, Duccio Piccardi, Silvia Calamai
• Regione Toscana (2019-2021)
• Web infrastructure hosting regional oral archives
  • Digital philology, metadata, audio restoration
• Case study: Caterina Bueno ethnomusicological archive
  • 476 analogue carriers
  • 700+ hours of recordings
  • Field recordings, live events, rehearsals...

“[ORAL HISTORY], LIKE ANY OTHER [OCCUPATION], HARBORS A POSSIBILITY OF INFlicting REAL OR IMAGINED INJURY AND WRONGS UPON OTHERS, AND THOSE THINGS USUALLY... RESULT IN LITIGATION”
(WELCH IN NEUENSCHWANDER, 2014: XIII)
COPYRIGHT ISSUES
Compensation for Subsequent Uses

- Inalienable right in any circumstances, including non-profit

- Quantification of compensation for non-profit repurposing is problematic

- The streaming of an online archive should not be considered as a "public" repurposing

- Compensable "subsequent uses" should benefit an indeterminate/fairly large number of listeners

Low-level risk, BUT never include explicit surrenders to compensation, and authorized access is preferable
CONCLUSIONS

Performer rights
  Reproduction rights
  Diffusion rights

  Analogue reasoning

Informed consent forms
Licenses
THANKS FOR YOUR ATTENTION!
Breakout Rooms

Breakout rooms for the topics discussed today will be available on Day 3, Wednesday 29 September, from 13:25 to 13:45.