How can CLARIN archive and curate my resources?

Christoph Draxler
draxler@phonetik.uni-muenchen.de
Outline

- Relevant resources
- CLARIN infrastructure
  - European Research Infrastructure Consortium
  - National CLARIN infrastructures
- Data Seal of Approval
  - Data producer
  - Repository
  - Data consumer
- Showcase AsiCa
Relevant resources

- CLARIN serves the humanities and social sciences
  - with a focus on linguistic resources
  - i.e. speech and text data
  - plus image, video, sensor data
    - gestures, motion capture, mimics, physiological, OCR, geo-referenced data…
  - plus domain-specific processing
- Co-existence with other infrastructures
Current CLARIN resources

- **Text corpora, e.g.**
  - historical corpora
  - linguistically annotated corpora
- **Spoken language corpora, e.g.**
  - speech technology corpora
  - phonetics corpora
  - humanities, social sciences, psychology, medicine
- **Tools**
  - processing tools, workflows, toolchains
  - exchange formats
CLARIN infrastructure

- Humanities now recognized as a major source of data
  - cf. physics, astronomy
- Infrastructures needed to archive and provide access to data
- CLARIN now a ERIC
  - European central CLARIN organisation
  - national CLARIN infrastructures
Why CLARIN? Which archive?

• CLARIN requires substantial financial commitment
  • national bodies
  • stable institutions
  • precautions against failure

• Select a CLARIN archive by its service
  • Data Seal of Approval
Data Seal of Approvement

- 16 guidelines for the curation of resources
  - 3 for data producers
  - 10 for data repositories
  - 3 for data consumers
- Archives and centres must adhere to these guidelines
  - re-evaluation every two years
  - transparent process
- Public and non-trivial commitment
Towards sustainable and trusted data repositories

There are 16 guidelines that together determine whether your data repository qualifies for the Data Seal of Approval.
Data producer…

• deposits data in a repository
  • sufficient information to assess data quality
  • compliant with disciplinary and ethical norms
• provides data in recommended formats
• provides data together with its metadata
Repository…

- explicit *mission* in the area of digital archiving
- uses due diligence
  - compliance with legal regulations and contracts
  - including protection of human subjects
- documented processes and procedures for managing data storage
- long-term preservation of its digital assets
- explicit work flows across the data life cycle
Repository… (cont’d)

• responsibility for access and availability of the digital objects
• searchability and persistent reference to data
• integrity of the digital objects and metadata
• authenticity of the digital objects and metadata
• technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS
Data consumer...

- complies with repository’s access regulations
- conforms to any codes of conduct
  - generally accepted in the relevant sector
  - for the exchange and proper use of knowledge and information
- respects licence terms of the repository
Example: CLARIN Center BAS


The repository is not a legal entity on its own but is part of the Institute of Phonetics and Speech Processing which is not a legal entity on its own but part of the Ludwig Maximilian University Munich. The legal status of the Ludwig Maximilian University is “Körperschaft des Öffentlichen Rechts“. The repository is funded by the Institute of Phonetics and Speech Processing. The repository has agreements with external depositors about the right to archive the data. The depositors themselves are responsible for compliance with any legal regulations in the area where the data is collected. The repository enables the depositors to restrict access to their resources at various levels. Distributed copies elsewhere may not be made available to third parties.

Online template contract, code of conduct and terms of usage:
- [http://www.phonetik.uni-muenchen.de/Bas/BasTemplateContract.pdf](http://www.phonetik.uni-muenchen.de/Bas/BasTemplateContract.pdf)
- [http://www.phonetik.uni-muenchen.de/Bas/BasTemplateInformedConsent_en.pdf](http://www.phonetik.uni-muenchen.de/Bas/BasTemplateInformedConsent_en.pdf)
- [https://www.phonetik.uni-muenchen.de/Bas/BasTermsOfUsage_eng.pdf](https://www.phonetik.uni-muenchen.de/Bas/BasTermsOfUsage_eng.pdf)
10. The data repository enables the users to utilize the research data and refer to them.

Minimum Required Statement of Compliance:
- Theoretical: We have a theoretical concept.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

The repository provides various ways of utilizing the archived data via online tools as well as by downloading the data in formats commonly used by the research communities. For very large resources where online access is not (yet) technically feasible we also provide the possibility to distribute resources on standard media (such as DVD-R and/or hard discs). An advanced metadata search utility is provided, as well as a simple search tool for textual content. All metadata can be harvested via the OAI-PMH protocol. Unique persistent identifiers according to the Handle system are provided for each corpus and each session within the corpora.

Bas repository:
http://clarin.phonetik.uni-muenchen.de/BASRepository/index.php

BAS OAI-PMH endpoint:
https://webapp.phonetik.uni-muenchen.de/BASSRU/

sample query:
https://webapp.phonetik.uni-muenchen.de/BASSRU/?version=1.2&operation=searchRetrieve&query=Gott
Choosing an Archive

- Check out existing archives, e.g. via DSA
  - accepted in my field?
  - sufficient domain knowledge?
  - terms of usage?
- Contact responsible person
  - found on repository pages
  - negotiate terms of distribution
- Iterate through curation steps
  - metadata preparation, data upload, legal issues…
- Publish your resource via the centre
Showcase AsiCa

- Atlante Sintattico della Calabria
  - Dept. of Romance Linguistics, Prof. Krefeld (LMU)
- Speech recordings of Calabrese dialect
  - read speech
  - interviews
- 60 speakers, controlled demographics
  - 30 with long-term residence in Germany
- Annotated speech
  - orthographic transcription
  - syntactic events
• Metadata in Excel and SQL
  • easy to create and maintain
• CMDI metadata
  • automatic generation for session items
  • manually for 1 corpus item
AsiCa in BAS Repository

clarin.phonetik.uni-muenchen.de/BASRepository
Summary

• Archiving resources requires long-term commitment
  • think in centuries, not projects
  • your data will outlive your project
• Adhere to archiving standards from the start
  • CMDI for meta-data, de facto formats for data
• Accept professional support
  • set aside money for this
  funding agencies will make you do so anyway!