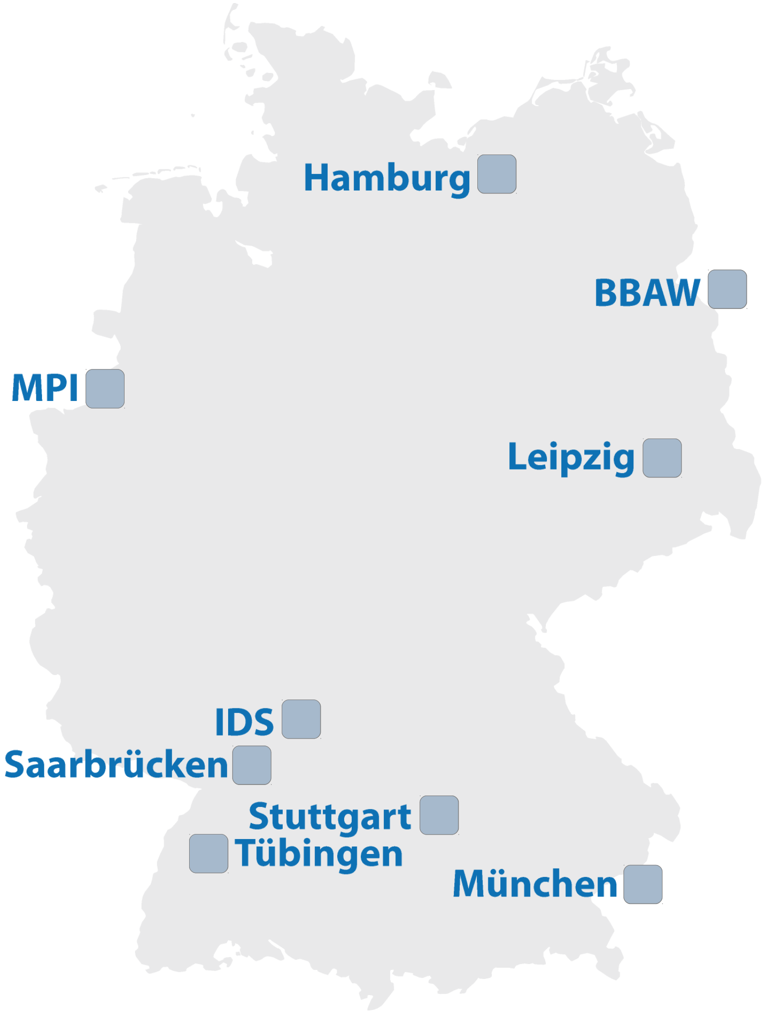
CLARIN Data Repositories

Florian Schiel

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Repositories of the CLARIN-D Infrastructure



CLARIN

- Federated union of persistent speech data providers
- Common file and metadata standards (DC,OLAC, CMDI), PIDs
- Common quality standards (repository software, general quality assessments)
- Search across data repositories (FCS)
- European infrastructure (central services, monitoring, center assessment)

-> *www.clarin.eu*

Repository Bavarian Archive for Speech Signals, Munich

BAS Repository

- specialized in speech- and multi-media corpora
- AAI authentification ('single sign on')
- download of corpora or sessions or mixed corpora extracts
- metadata search within BAS
- internal structure: center -> corpus -> session
- export of metadata to harvesters via OAI-PMH standard (e.g. OLAC)
- content search across CLARIN centers (Federated Content Search)
- Persistent Identifiers (PID)

Repository Bavarian Archive for Speech Signals, Munich

Demo:
https://clarin.phonetik.uni-muenchen.de/BASRepository

- *show CLARIN help desk access*
- *show single-sign-on login via AAI*
- *show general structure: center -> corpus -> session -> media/annotations*
- *show example metadata corpus: Dublin Core (DC), CMDI*
- *show example metadata session: CMDI*
- *show metadata search:*
  German monologues of native German female speakers aged 27
- *show content search via CLARIN FCS*
  http://weblicht.sfs.uni-tuebingen.de/Aggregator/

BAS Corpus Ingest

- ingest = *incorporating/updating a corpus into the repository*
- required input from corpus provider -> BAS :
  + (standard) media files
  + annotation files
  + documentation package (ZIP)
  + metadata (e.g. CSV tables, DC, OLAC, CMDI)
- ingest process :
  + upload raw corpus
  + BAS validation
  + create BAS conform CMDIs that point to corpus
  + ingest : checks, PID assignment, version control etc.
- *demo: /vdata/CLARIN/Metadata/Corpora/test/ALCtest/v1/*

Advantages of a CLARIN Repository

- long-term archival storage
  including format conversions etc.
- wide dissemination to prospective users
  via repository portal, via harvesting organisations ...
- quality control (standards, documentation ...)
- persistent citation via PID
  e.g. hdl.handle.net/11858/00-1779-0000-001A-1A81-E
- version control
  e.g. adding new annotations, bug fixes, extensions ...
- assistance
  e.g. CMDI creation, validation