

# CLARIN-D-Ressourcen:

Was kann ich nutzen und wie geht das?



Axel Herold

17. / 18. November 2014,  
Textkorpora in Infrastrukturen für die Geistes- und  
Sozialwissenschaften, CLARIN-D-Workshop

## CLARIN —

### Common Language Resources and Technology Infrastructure

- ▶ Sprachressourcen und -werkzeuge für Geistes- und Sozialwissenschaftlerinnen erschließen
- ▶ Integration vieler verschiedenartiger linguistischer Ressourcen und Werkzeuge
- ▶ **CLARIN-D**: BAS, BBAW, IDS Mannheim, MPI Nijmegen, Uni Hamburg, Uni Leipzig, Uni Saarland, Uni Stuttgart, Uni Tübingen (RZ Garching, FZ Jülich)
- ▶ **CLARIN ERIC** (europäischer Rahmen): CLARIN-AT, LINDAT-CLARIN (CZ), CLARIN-D, DLU, CLARIN-DK, CELR (EE), CLARIN-NL, CLARIN-PL, SWE-CLARIN, CLARIN-BG, CLARIN-LT (SWE-CLARIN, CLARINO)

## Vision: (europaweit) vernetzte Forschungsinfrastruktur

- ▶ Auffindbarkeit von Ressourcen
- ▶ Erschließung für verschiedene Nutzergruppen  
(mehr dazu: Boehlke)
- ▶ Persistenz
- ▶ Interoperabilität
- ▶ offene Schnittstellen
- ▶ Zugriffskontrolle
- ▶ Hosting-Services
- ▶ Dokumentation, Schulung, Hilfe

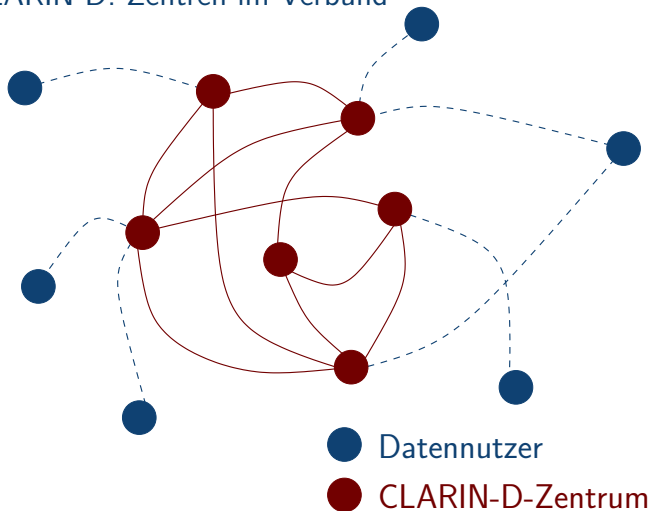
## CLARIN-D: Zentren im Verbund

- ▶ Repositorien  
(Primär- und Metadaten, Versionierung, Archivierung,  
mehr dazu: Zimmer)
- ▶ Endpunkte für Korpusrecherchen
- ▶ sprachtechnologische Webservices  
(mehr dazu: Trippel)
- ▶ Identity-Provider
- ▶ ...

## Zentrale Dienste

- ▶ Virtual Language Observatory (VLO)
- ▶ Virtual Collection Registry (VCR)
- ▶ Federated Content Search (FCS)
- ▶ WebLicht Webservice-GUI
- ▶ ...

## CLARIN-D: Zentren im Verbund



## Einige mögliche Szenarios

- ▶ Metadatensuche, Ressource(n) lokal nutzen  
(VLO → konkretes Repository)

## Einige mögliche Szenarios

- ▶ Metadatensuche, Ressource(n) lokal nutzen  
(VLO → konkretes Repository)
- ▶ linguistische Analyse/Annotation eigener Daten  
(WebLicht)



## Einige mögliche Szenarios

- ▶ Metadatensuche, Ressource(n) lokal nutzen  
(VLO → konkretes Repository)
- ▶ linguistische Analyse/Annotation eigener Daten  
(WebLicht)
- ▶ Archivierung eigener Forschungs(primär)daten  
(Repository mit entsprechendem Profil, VCR)

## Einige mögliche Szenarios

- ▶ Metadatensuche, Ressource(n) lokal nutzen (VLO → konkretes Repositorium)
- ▶ linguistische Analyse/Annotation eigener Daten (WebLicht)
- ▶ Archivierung eigener Forschungs(primär)daten (Repositorium mit entsprechendem Profil, VCR)
- ▶ Bereitstellen eigener Ressourcen und Dienste (CLARIN-Standards)
- ▶ ...

## Generell: Nutzung der Infrastruktur

- ▶ über Benutzer-Schnittstellen (GUIs)
- ▶ über standardisierte technische Schnittstellen

## CLARIN-Ressourcen

- ▶ Überblick: Virtual Language Observatory
- ▶ facettierte Suche
- ▶ ca. 600 000 Ressourcen
- ▶ **Achtung:** Granularität
- ▶ stets aktuelle Daten durch Repositorien bereitgestellt
- ▶ (keine prinzipiellen Formatbeschränkungen)
- ▶ (mehr dazu: van Uytvanck)

## Metadateninfrastruktur: CMDI

- ▶ Component Metadata Infrastrukture
- ▶ keine prinzipiellen Einschränkungen der Beschreibungsdomäne
- ▶ modular, adaptierbar (Profile, Komponenten)
- ▶ Interoperabilität durch semantische Annotation (CLARIN Concept Registry, CCR)
- ▶ momentan: ISO-Standardisierung

## Datentypen

- ▶ Texte
- ▶ Wörterbücher
- ▶ Tonaufnahmen
- ▶ Videoaufnahmen
- ▶ Annotationen
- ▶ Datensätze
- ▶ Webservices, Anwendungen
- ▶ (keine prinzipiellen Typbeschränkungen)

## Unterstützte Primärdatenformate

- ▶ Unterstützungsgrad abhängig von Infrastrukturkomponenten
- ▶ Plain-Text Formate (viele Webservices)
- ▶ nativ: TCF (Text Corpus Format, auch für Wörterbücher)
- ▶ ausgewählte TEI-Dialekte (z. B. DTABf)
- ▶ anwendungsspezifische Audio- / Videoformate
- ▶ vereinzelt Unterstützung für spezialisierte Datenformate (z. B. Wörterbücher)
- ▶ (mehr zu einzelnen Themen: Haaf / Jurish)

**Strategie:** Bereitstellen von Konvertern

## CLARIN-D-Benutzerhandbuch

- ▶ „CLARIN-D in a nutshell“ – gute Einstiegsmöglichkeit und Hintergrundinformationen
  - ▶ grundlegende Eigenschaften und Modellierungsprinzipien für Ressourcen und Werkzeuge
  - ▶ spezifische Eigenschaften und Anforderungen
- ▶ <http://de.clarin.eu/de/sprachressourcen/benutzerhandbuch.html>

## Weitere Hilfe

- ▶ Helpdesk
- ▶ Beratung durch Zentren
- ▶ (mehr dazu: Lehmberg)

## Offene Punkte

- ▶ Weiterentwicklung des Datenangebots  
(in Abstimmung mit den Facharbeitsgruppen)
- ▶ nativer Umgang mit verbreiteten Primär(text)formaten  
(bspw. TEI-Integrator, TEI-Dokumente in WebLicht nutzen)
- ▶ breitere Erschließung von nicht textbasierten Daten  
(Ton-, Filmaufnahmen)
- ▶ Harmonisierung der Metadaten



# Danke! Fragen?

<http://de.clarin.eu/>

<http://www.clarin.eu/>