CLARIN(-D)

Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften

Thomas Eckart
Abt. Automatische Sprachverarbeitung
Institut für Informatik, Universität Leipzig
teckart@informatik.uni-leipzig.de

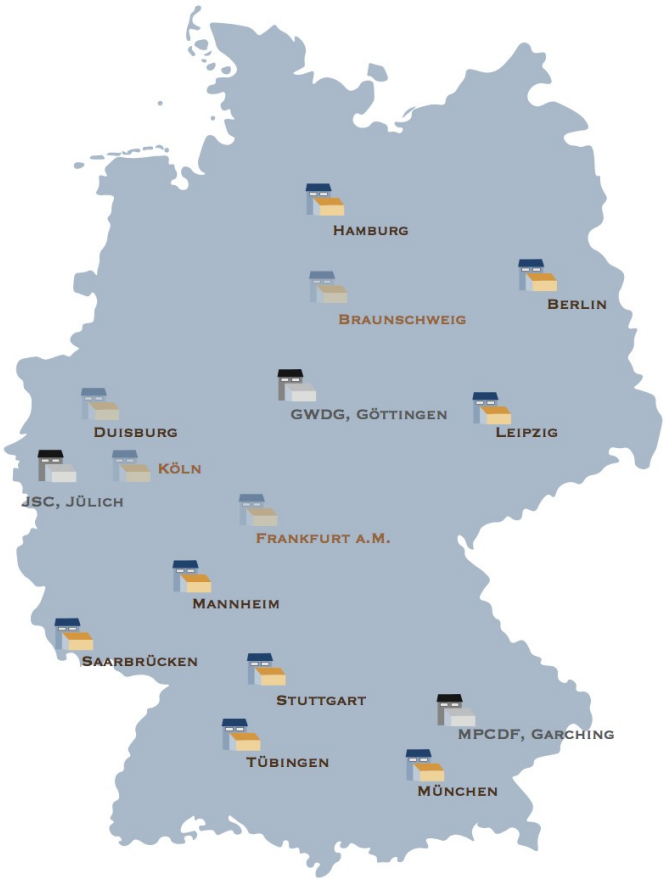
UNIVERSITÄT LEIPZIG

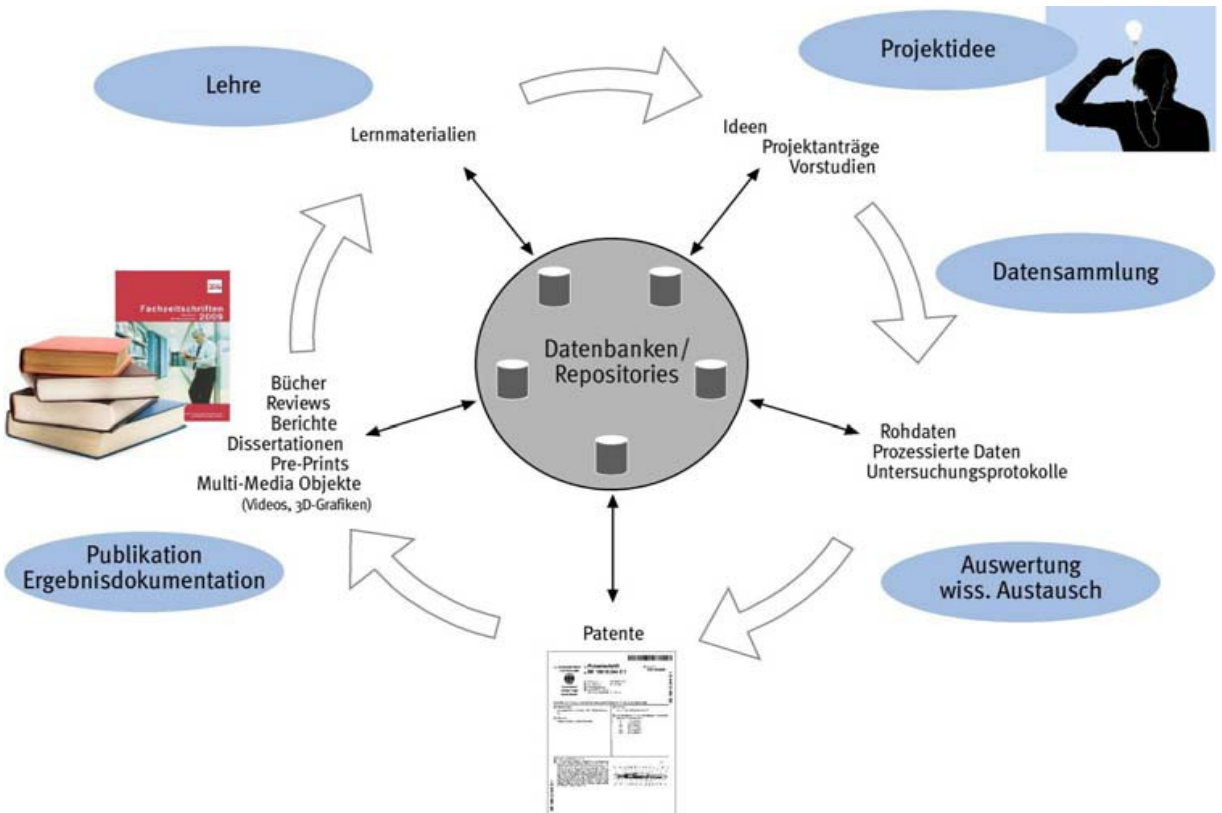
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

CLARIN-D – Web- und zentrenbasierte Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften, ESFRI

Sprachdaten-, -tools und services:

- In einer integrierten, interoperable und skalierbaren Infrastruktur
- Gefördert durch das Bundesministerium für Bildung und Forschung (BMBF)
- Aktuelle Projektphase: bis 31.12.2020
- Web: http://de.clarin.eu
- Laufende und zukünftige Anträge:
  - Integration mit DARIAH-DE (CLARIAH) bis 2021
  - NFDI (Text+)



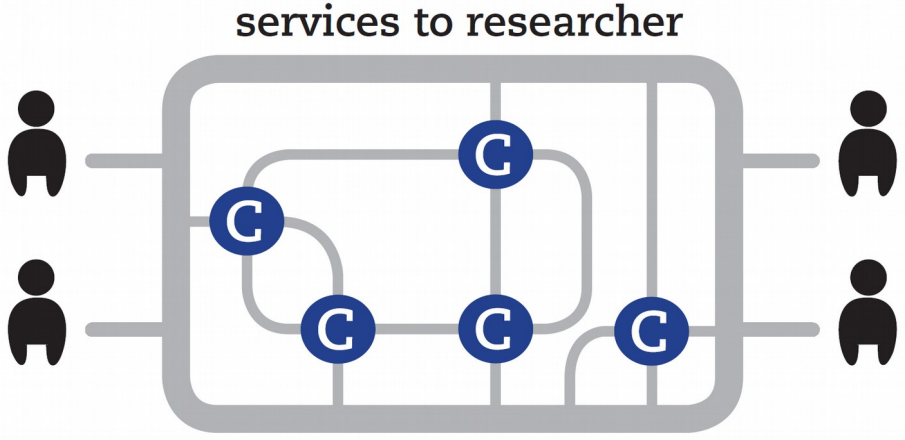

Lehre
Projektidee
Lernmaterialien
Ideen
Projektanträge
Vorstudien
Datensammlung
Datenbanken/
Repositories
Bücher
Reviews
Berichte
Dissertationen
Pre-Prints
Multi-Media Objekte
(Videos, 3D-Grafiken)
Rohdaten
Prozessierte Daten
Untersuchungsprotokolle
Publikation
Ergebnisdokumentation
Auswertung
wiss. Austausch
Patente

Architektur

Verteilte Infrastruktur mit

- Standardisierten Schnittstellen (für Metadaten, Daten und Werkzeuge/Services),
- Verarbeitungsschichten und
- Kommunikationsprotokollen

Nutzer sowohl Anbieter und Konsumenten



CLARIN-D

- Zentren (B):
  - Berlin: BBAW
  - Mannheim: IDS
  - Universität Hamburg
  - Universität Leipzig
  - Universität München (LMU)
  - Universität des Saarlands
  - Universität Stuttgart
  - Universität Tübingen
- Rechenzentren (E):
  - Garching: RZG
  - Göttingen: GWDG
  - Jülich: RZJ

Ansprechpartner

Bitte beschreiben Sie ihre Daten mit den bestmöglichen Kategorien.

Welche Art von sprachwissenschaftlichen Daten liegen vor?

- ☐ Gesprochene Sprache
- ☑ Geschriebene Sprache
- ☐ Multimodale Sprachdaten
- ☐ Gebärdensprache

Um welche Sprache(n) handelt es sich?

- ☐ Mehrsprachige Daten
- ☑ Deutsche Sprache
- ☐ Andere Sprachen (nicht Deutsch)
- ☐ Minderheitensprache
- ☐ Bedrohte Sprache
- ☐ Historische Sprache
- ☐ Gegenwartssprache

Sind die Daten öffentlich (oder unter Lizenz) ?

● Ja ● Nein

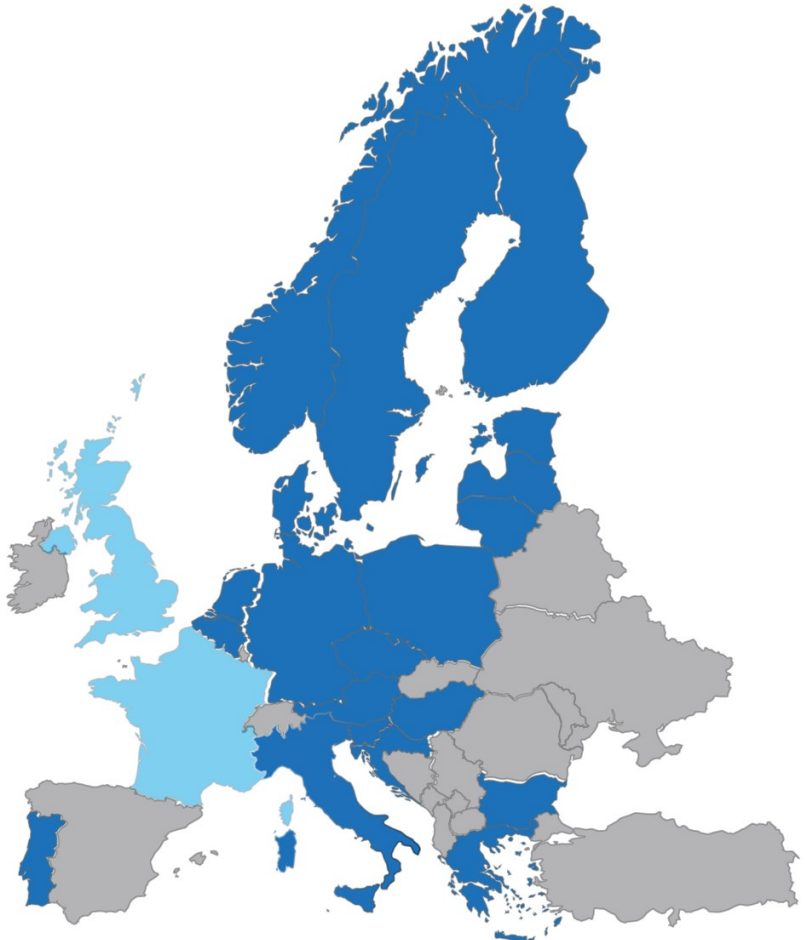
Welcher Ressourcentyp liegt vor?

- ☑ Lexikon
- ☐ Korpus
- ☐ Baumbank
- ☐ Digitale Editionen
- ☐ Experimentaldaten
- ☐ Sprachtechnologische Daten
- ☐ Andere Textdaten

Es gibt 4 Zentrum/Zentren zur Archivierung ihrer Daten! Bitte kontaktieren Sie unten aufgeführten Clarin-D Zentren:

- [1] Berlin-Brandenburg Academy of Sciences and Humanities, Alexander geyken@bbaw.de
- [2] Eberhard Karls Universität Tübingen, Marie Hinrichs, marie.hinrichs@tuebingen.de
- [3] ASV Leipzig, Prof. Dr. Gerhard Heyer, heyer@informatik.uni-leipzig.de
- [4] Liaison Contact, Dr. Thorsten Trippel, thorsten.trippel@uni-tuebingen.

- Mitglieder:
  - Austria
  - Belgium
  - Bulgaria
  - Croatia
  - Cyprus
  - Czech Republic
  - Denmark
  - Estonia
  - Finland
  - Germany
  - Greece
  - Hungary
  - Italy
  - Latvia
  - Lithuania
  - Netherlands
  - Norway
  - Poland
  - Portugal
  - Slovenia
  - Sweden
- Kandidaten (Observer):
  - Iceland
  - France
  - South Africa
  - United Kingdom



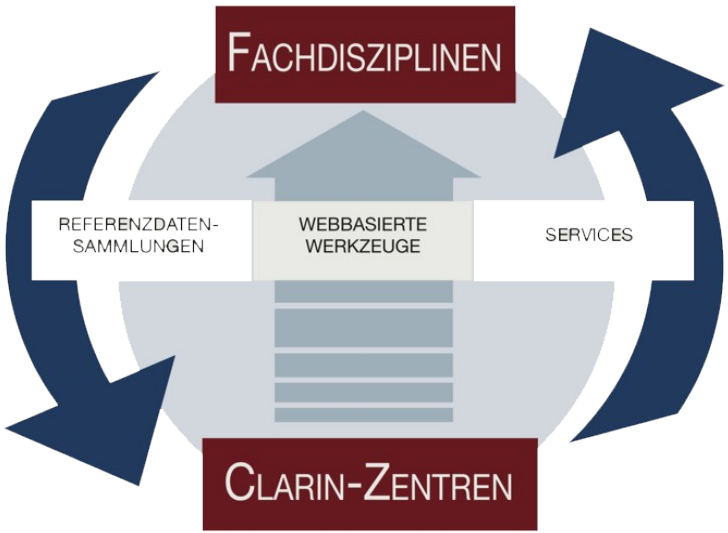
Einbindung der Fachwissenschaften

Facharbeitsgruppen (F-AG)

Enge Zusammenarbeit mit Fachwissenschaftlern:

- F1: Deutsche Philologie
- F2: Andere Philologien
- F3: Linguistische Feldforschung, Ethnologie, Typologie
- F4: Menschliche Sprachverarbeitung: Psycholinguistik, Kognitionspsychologie
- F5: Gesprochene Sprache und andere Modalitäten
- F6: Angewandte Sprachwissenschaft, Computerlinguistik
- F7: Inhaltsanalytische Methoden in den Sozialwissenschaften
- F8: Geschichtswissenschaften

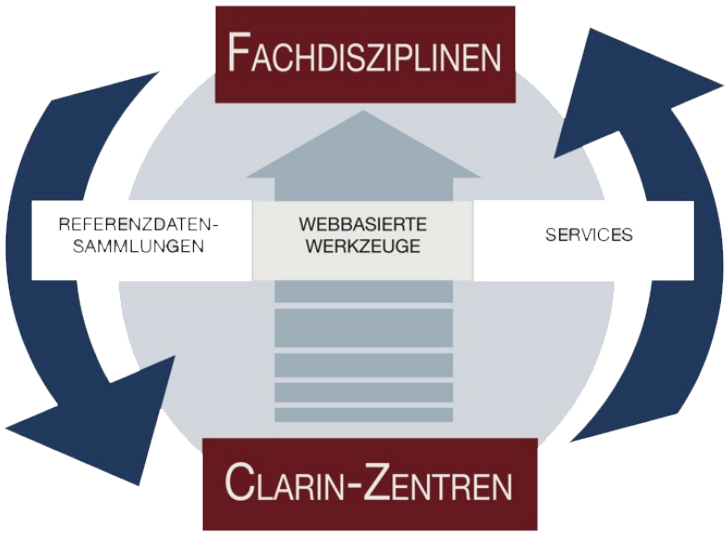
Insgesamt: ~200 Mitglieder



Facharbeitsgruppen (F-AG)

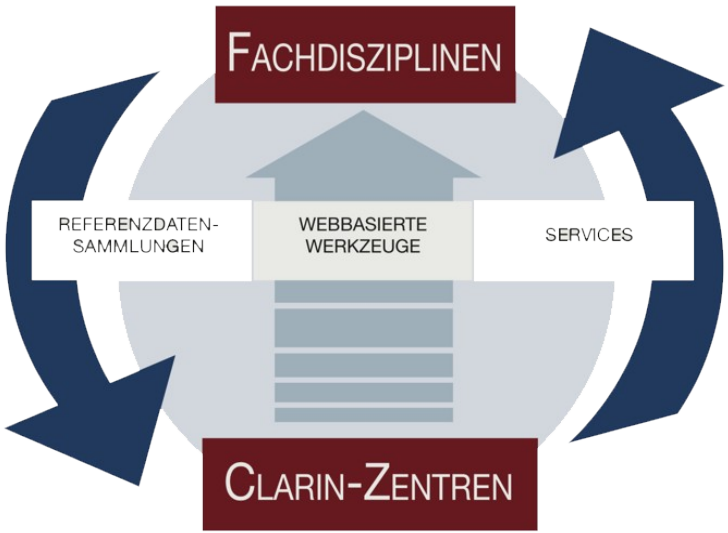
- Kurationsprojekte als „Starthilfe“

- Beispiele:
  - F2: „CLARIN-Integration des Old Bailey Korpus“ (Gießen/Saarland)
  - F4: „Implementierung einer Plattform für Open Science / Reproducible Research für Psycholinguistik und Kognitionspsychologie“ (Potsdam/Leipzig)
  - F7: „Semantische Annotation für Digital Humanities“ (Heidelberg/Darmstadt)



Weitere Unterstützung

- Ressourcen in Zentren
  - für Auftragsarbeiten (Usecases), z.B.
    - Weboberfläche für Korpusvergleich (Leipzig)
    - Crawling-Webanwendung (Leipzig)
  - Beratung
  - Teilnahme am zentralen CLARIN-Helpdesk



Vertrauenswürdigkeit - Qualitätsstandards

Zertifizierung der Compliance mit verschiedenen Qualitätsstandards

Core Trust Seal (CTS / DSA), 3 Jahre

- Zertifizierung von Datenrepositories basierend auf allgemeinem Anforderungskatalog
  - Allgemeine Anforderungen an vertrauenswürdiges, langfristiges Hosting von Ressourcen
  - Organisatorische, technische, finanzielle und rechtliche Aspekte
- Nonprofit-Organisation gewidmet nachhaltiger und vertrauenswürdiger Dateninfrastrukturen (unter Schirmherrschaft der Research Data Alliance)
- Vorteile für Nutzer der Infrastruktur:
  - Externe Checks des Datenmanagements
  - Compliance mit etablierten Prozeduren
  - Transparenz (interner) Prozesse im Repositorium

Vertrauenswürdigkeit - Qualitätsstandards

Zertifizierung der Compliance mit verschiedenen Qualitätsstandards

CLARIN-Zentrenzertifizierung, 3 Jahre

- Diverse CLARIN-spezifische Anforderungen
  - Verwendung von Persistent Identifiers, Federated Identity Management, Bereitstellung von Schnittstellen, Auslieferung und Mindeststandards von Metadaten, etc.

CLARIN B CENTRE

Vielen Dank!

Kommentare? Fragen?