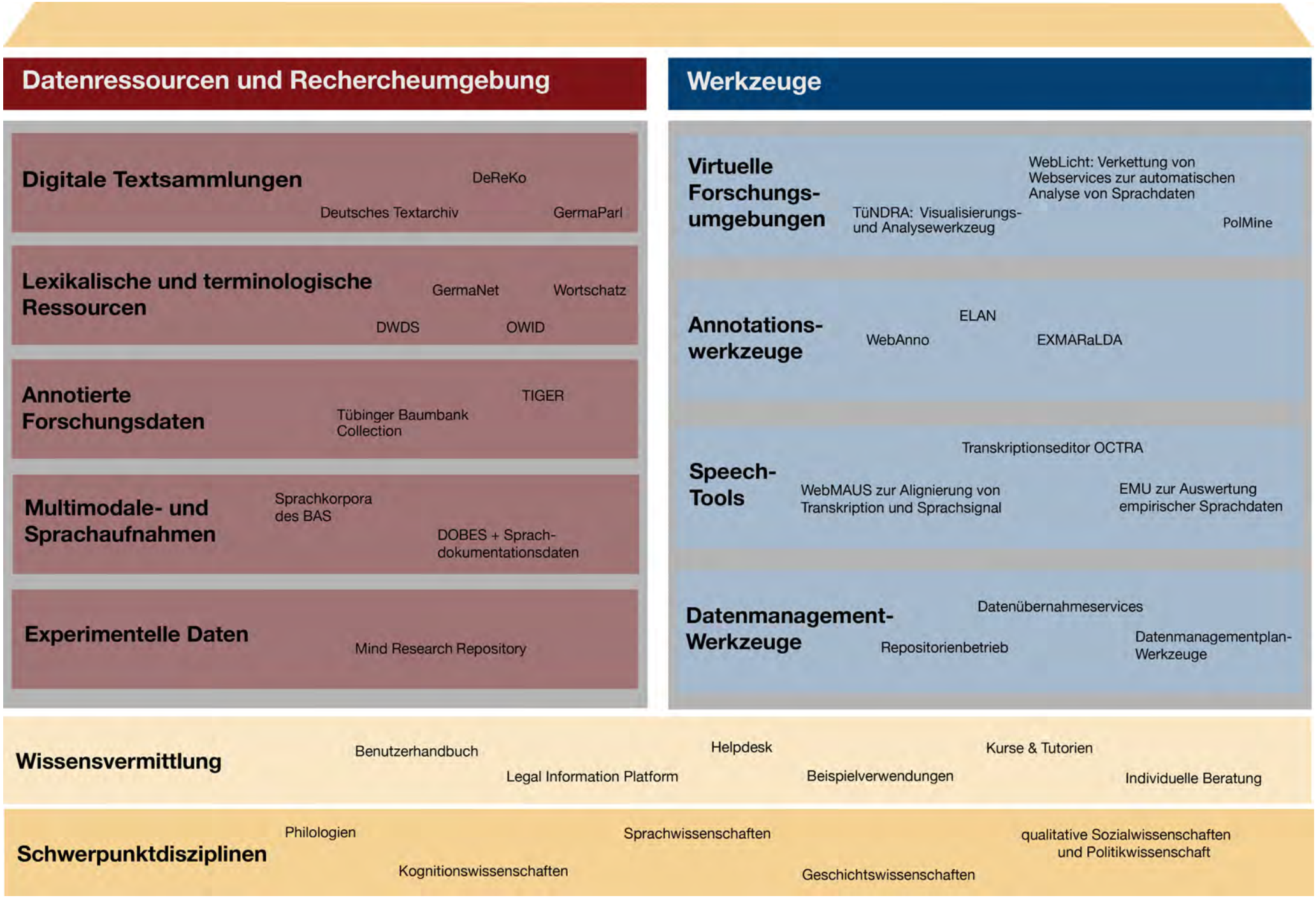
CLARIN-D: Digitale Forschungsinfrastruktur in den Geisteswissenschaften

Prof. Dr. Erhard Hinrichs
Eberhard Karls Universität Tübingen

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung


Datenressourcen und Rechercheumgebung
Digitale Textsammlungen
DeReKo
Deutsches Textarchiv
GermaParl
Lexikalische und terminologische Ressourcen
GermaNet
Wortschatz
DWDS
OWID
Annotierte Forschungsdaten
TIGER
Tübinger Baumbank Collection
Multimodale- und Sprachaufnahmen
Sprachkorpora des BAS
DOBES + Sprach-dokumentationsdaten
Experimentelle Daten
Mind Research Repository
Werkzeuge
Virtuelle Forschungs-umgebungen
WebLicht: Verkettung von Webservices zur automatischen Analyse von Sprachdaten
TüNDRA: Visualisierungs- und Analysewerkzeug
PolMine
Annotations-werkzeuge
ELAN
WebAnno
EXMARaLDA
Speech-Tools
Transkriptionseditor OCTRA
WebMAUS zur Alignierung von Transkription und Sprachsignal
EMU zur Auswertung empirischer Sprachdaten
Datenmanagement-Werkzeuge
Datenübernahmeservices
Repositorienbetrieb
Datenmanagementplan-Werkzeuge
Wissensvermittlung
Benutzerhandbuch
Helpdesk
Kurse & Tutorien
Legal Information Platform
Beispielverwendungen
Individuelle Beratung
Schwerpunktdisziplinen
Philologien
Sprachwissenschaften
qualitative Sozialwissenschaften und Politikwissenschaft
Kognitionswissenschaften
Geschichtswissenschaften

Aufbewahren von Sprachressourcen

- Repositorienbetrieb
  - Zertifizierte Archivsysteme an allen CLARIN-D-Zentren
  - CLARIN-externe und interne Evaluierung
- Datenmanagementplan Werkzeuge
  - Erstellung von Datenmanagementplänen bei Projektbeantragungen
  - Kooperation beim Datenmanagement über den gesamten Projektzyklus
- Datenübernahmeservices
  - OAIS-unterstützte Archivierungsverfahren
  - Erstellung von *submission information packages* (SIP-Archivierungspaketen) im Dialog zwischen Datengebern und –Archiven
  - Basiert auf Datenhinterlegungsverträgen

Datenressourcen und

Digitale Textsammlunger

Lexikalische und termino Ressourcen

Annotierte Forschungsdaten

Multimodale- und Sprachaufnahmen

Experimentelle Daten

Wissensvermittlung

Schwerpunktdisziplinen

Facharbeitsgruppen für unterschiedliche Disziplinen

> 200 Forschende in ganz Deutschland in 8 Facharbeitsgruppen

- Philologien (Amerikanistik, Anglistik, Germanistik, Klassische Philologie, Romanistik, Slavistik)
- Sprachwissenschaften (inkl. Computerlinguistik, Sprachtechnologie, Phonetik und allgemeine Linguistik)
- Kognitionswissenschaften
- Geschichtswissenschaften
- qualitative Sozialwissenschaft und Politikwissenschaft
- Kulturwissenschaften (inkl. Anthropologie und Sprachdokumentation)

CLARIN-D: offener Zentrenverbund

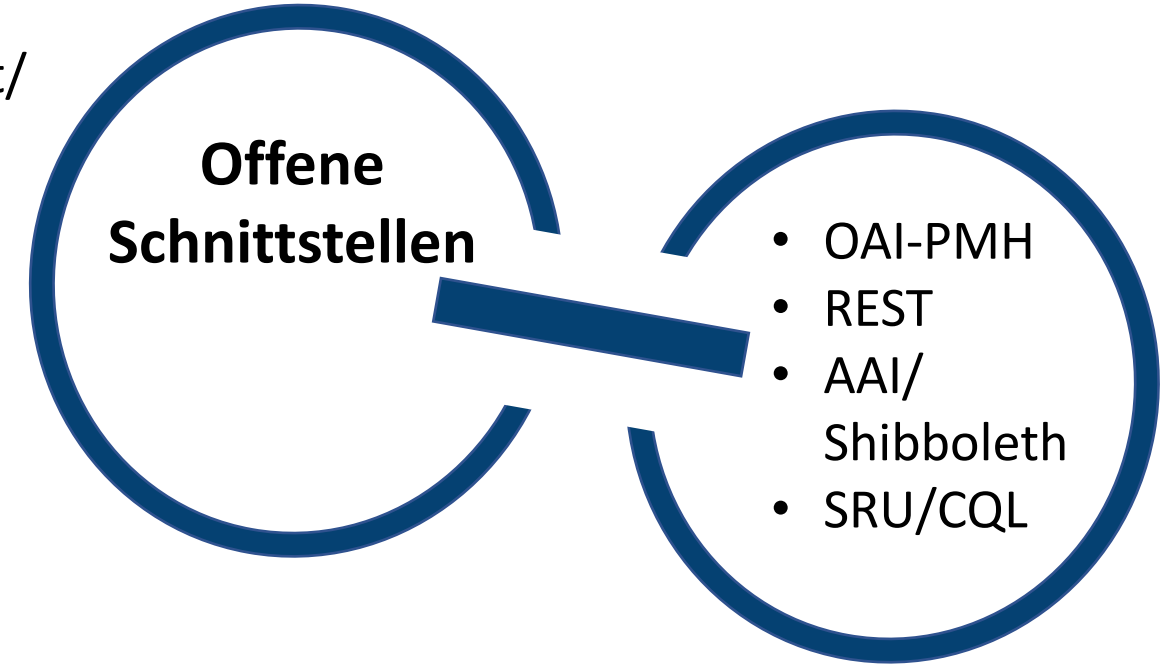
- zertifizierte Datenzentren
  - gemeinsame technischen Infrastruktur und Standards
  - Datenmanagementangebote an Dritte
- angebunden an europäische Infrastruktur
- offene Verbundsarchitektur

- Zentren im Konsortium
- Assoziierte Zentren
- Kooperierende Rechenzentren

CLARIN-D Architektur

Zentrenangebote

- Datenmanagement/ Repositorien
- Webservices
- Webanwendung
- Beratung
- Hilfe



Föderierte Werkzeuge

- Federated Content Search (FCS)
- WebLicht Chainer
- Virtual Language Obervatory (VLO)
- Virtual Collection Registry

Normen, Standards und andere Best Practices

Einsatz von Standards

- Metadaten:
  - **ISO 24622-1, ISO/DIS 24622-2**
  - Language resource management -- Component Metadata Infrastructure (CMDI)
  - Umsetzung durch alle CLARIN Zentren in Katalogmetadaten passend zum Datentyp
- PID:
  - **ISO 24619**
  - Language resource management -- Persistent identification and sustainable access (PISA)
  - Umsetzung durch Handle-System (empfohlen), DOI, URN

Einsatz von Standards: Spezifische Datentypen

- **ISO 24610-X:** Merkmalsstrukturen „Language resource management -- Feature structures“
- **ISO 24612:** Annotationen „Language resource management -- Linguistic annotation framework (LAF)“
- **ISO 24613**: Lexikalische Ressourcen „Language resource management - Lexical markup framework (LMF)“
- **ISO 24615**: Syntaktische Annotationen „Language resource management -- Syntactic annotation framework (SynAF)“
- **ISO 24623-X**: Auswertung von Korpora: „Language resource management -- Corpus query lingua franca (CQLF)“
- **ISO 24624**: Transkription gesprochener Sprache „Language resource management -- Transcription of spoken language“

Einsatz von Standards: Weitere Standards und Regeln

- Text Encoding Initiative (TEI)
  - DTA-Basisformat (DTABf): DFG-Empfehlung
- Encoding nach UTF-8
- Sprachkennzeichnung und Länder, etc.:
  - IETF RFC 5646 (für XML-Lang)
  - ISO 639 Sprachcodes (insbesondere ISO 639-3)
  - ISO 3166 Ländercodes
  - Normdaten, etwa GND, VIAF, ORCID, ...

Schnittstellen, Interoperabilität

- Internetstandards:
  - TCP/IP
  - HTTP
- Metadaten:
  - OAI-PMH
- Föderierte Suche:
  - SRU/CQL: Search/Retrieve via URL – Contextual Query Language (OASIS-Standard)
- Webservices:
  - Representational State Transfer (REST)

Voraussetzung für die Langzeitarchivierung der Daten

- Verankerung in Forschung und Lehre
  - Verwendung von Daten, Diensten und Technologien in der Ausbildung von Forschenden
  - Verwendung von Daten, Diensten und Technologien in Forschungsprojekten
  - Datenmanagementpläne im Antragsrahmen von Projekten und durchgeführt mit wissenschaftsgeleiteten Forschungsinfrastrukturanbietenden
- Institutionalisierung
- Verbund von Zentren: Fallback-Lösungen
- Einbindung in nationale und internationale Forschungsinfrastruktur-Netzwerke
  - Europa: Rechtsform ERIC European Research Infrastructure Initiative
  - Deutschland: derzeitige Diskussion der Architektur für nationale Verbünde
- Standardbasierte Aufbereitung und Repräsentation von Daten
- Dokumentierte, standardisierte Abläufe für die Datenverwaltung

Kurzer Überblick in Zahlen

- **1.677.412** Forschungsdatensätze (über europäischen Verbund)
- **> 170.000** Downloads aus deutschen Zentren jährlich
- **> 400** Webservices in WebLicht integriert
- **>5.800.000** Service Calls
- **>100** Annotationsvorhaben mit WebAnno
- **> 33.000** registrierte Endbenutzer
- **> 800** Lizenznehmer
- **462** kooperierende Projekte
- **2.421** Institutionen
- **35** beteiligte Fachdisziplinen

Stand Mai 2018