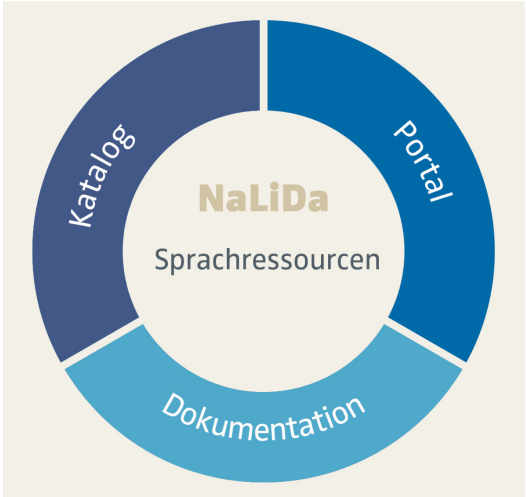
NaLiDa: Nachhaltigkeit linguistischer Daten

Erhard Hinrichs, Thorsten Trippel, Claus Zinn
Seminar für Sprachwissenschaft
Universität Tübingen

3. DINI/nestor-Workshop "Langzeitarchivierung von Forschungsdaten"
19. Juni 2015, ZIB in Berlin

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



ZENTRUM FÜR DATENVERARBEITUNG (ZDV)

UNIVERSITÄTSBIBLIOTHEK

Leitlinien der DFG

1.3.3 ***Forschungsdaten****, DFG Vordruck 60.100 - 09/14*

... Wenn im Rahmen des geplanten Sonderforschungsbereichs systematisch (Mess-)Daten generiert werden, die für die Nachnutzung geeignet sind, legen Sie bitte dar, welche Maßnahmen ergriffen wurden bzw. während der Laufzeit des Projektes ergriffen werden sollen, um die ***Daten nachhaltig zu sichern*** *und ggf.* ***für eine erneute Nutzung bereit zu stellen****. Bitte berücksichtigen Sie dabei auch – sofern vorhanden – die* ***existierenden Standards*** *und die* ***Angebote bestehender Datenrepositorien****. Stellen Sie bitte auch dar, durch* ***welche Einrichtungen*** *(Datenkurator, Rechenzentrum, Bibliothek, etc.)* ***welche Form von Unterstützung beim Daten- und Informationsmanagement*** *durch die am Sonderforschungsbereich beteiligten Institutionen geleistet werden soll.“*

Datenmanagement-Pläne werden auch von AHRC, EPSRC, ESRC, NSF, Wellcome Trust *etc.* erwartet

Projektüberblick

- digitale Archivierung **linguistischer** Forschungsprimärdaten
  - **Sammlung** von wissenschaftlichen Daten im sprachwissenschaftlichen Umfeld
  - Verwendung und Weiterentwicklung eines **Metadatenframeworks** zu ihrer Beschreibung
  - Einpflegen der Daten in ein **Repositorium**
  - Entwicklung von **Werkzeugen** zur Metadatenverarbeitung und zum Metadaten-basierten Zugriff
- **Referenzmodell**
  - technologischen Infrastruktur
  - standardisierte und qualitätssichernde Arbeitsabläufe
  - Integration in existierende universitäre Infrastruktur

Projektüberblick

- individuelle Beratungs- und Servicedienstleistungen
- Aufbau eines nationalen und internationalen Verbunds

Funding 3+2 Jahre, 2. Phase ab 2015

The project for Sustainability of Linguistic Data (NaLiDa) is funded by the German Research Foundation (DFG) in the program for Scientific Library Services and Information Systems (LIS).

DFG Deutsche Forschungsgemeinschaft

Sprachwissenschaftliche Forschungsprimärdaten

- Korpora (Sammlungen von Texten und Sprachaufnahmen)
  - TüBa-D/Z Baumbank (angereichert mit syntaktischen Struktur)
- Lexika (z.B. GermaNet: deutsche Version von WordNet)
- Grammatiken (z.B. HPSG Grammatik für Deutsch)
- Experimentaldaten (z.B. Eyetracking Daten zur Sprachperzeption)
- Software (z.B. Webservices zur Analyse von Sprachdaten, Statistiksoftware)
- Bilddaten zur Verarbeitung von Sprache im Gehirn (Magnetoenzephalographie & funktionelle Magnetresonanztomographie)

760.000 erfasste Ressourcen, davon über 60TSD deutsch-sprachige

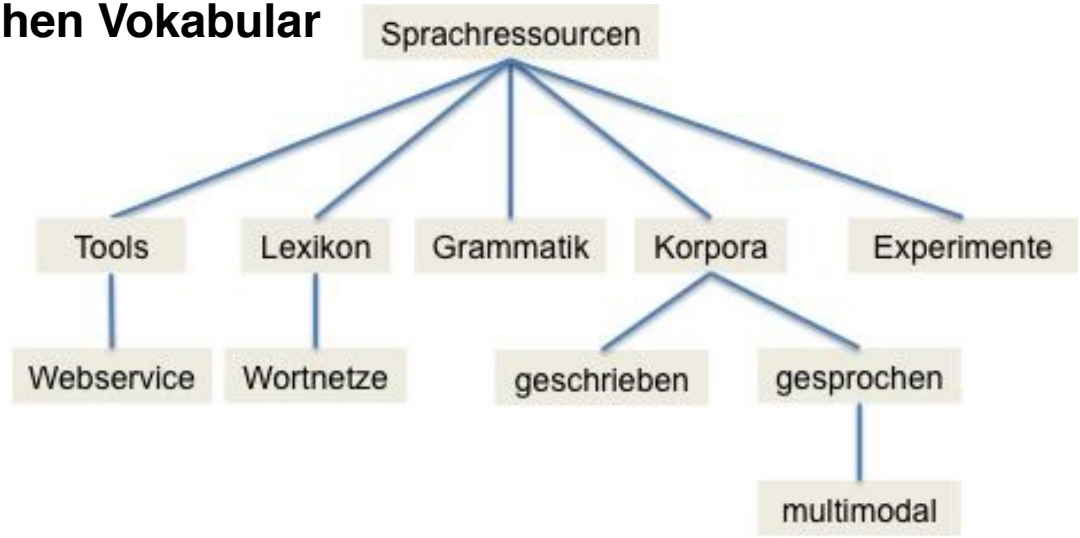
Datenformate

- unterschiedlichste Daten- und Informationsobjekte in Form von Text, Bild, Audio und Video
- Primärdaten häufig mithilfe diverser Software mit Zusatzinformationen angereichert
- Vielzahl von Datenformaten & Softwareprogrammen, die nur in definierten technischen Kontexten funktionieren
- Multimediaformate WAV, MP3 und MPEG2
- Formate R, SPSS, Excel, Praat und Text (OnEXP) für Experimentaldaten
- die textbasierten Formate PDF, Word, und XML
- die Dateiarchivformate Zip, GZip, Tar, und Rar

Datenherstellern wird empfohlen, proprietäre Formate (z.B. Word, SPSS) in nicht-proprietäre Formate (z.B. ODF, R) zu transformieren.

Beschreibung der Forschungsdaten mit Metadaten

- **Beschreibung mit fachwissenschaftlichen Vokabular**
- Hierarchie von Datentypen
- Jeder Typ hat eigenes Schema



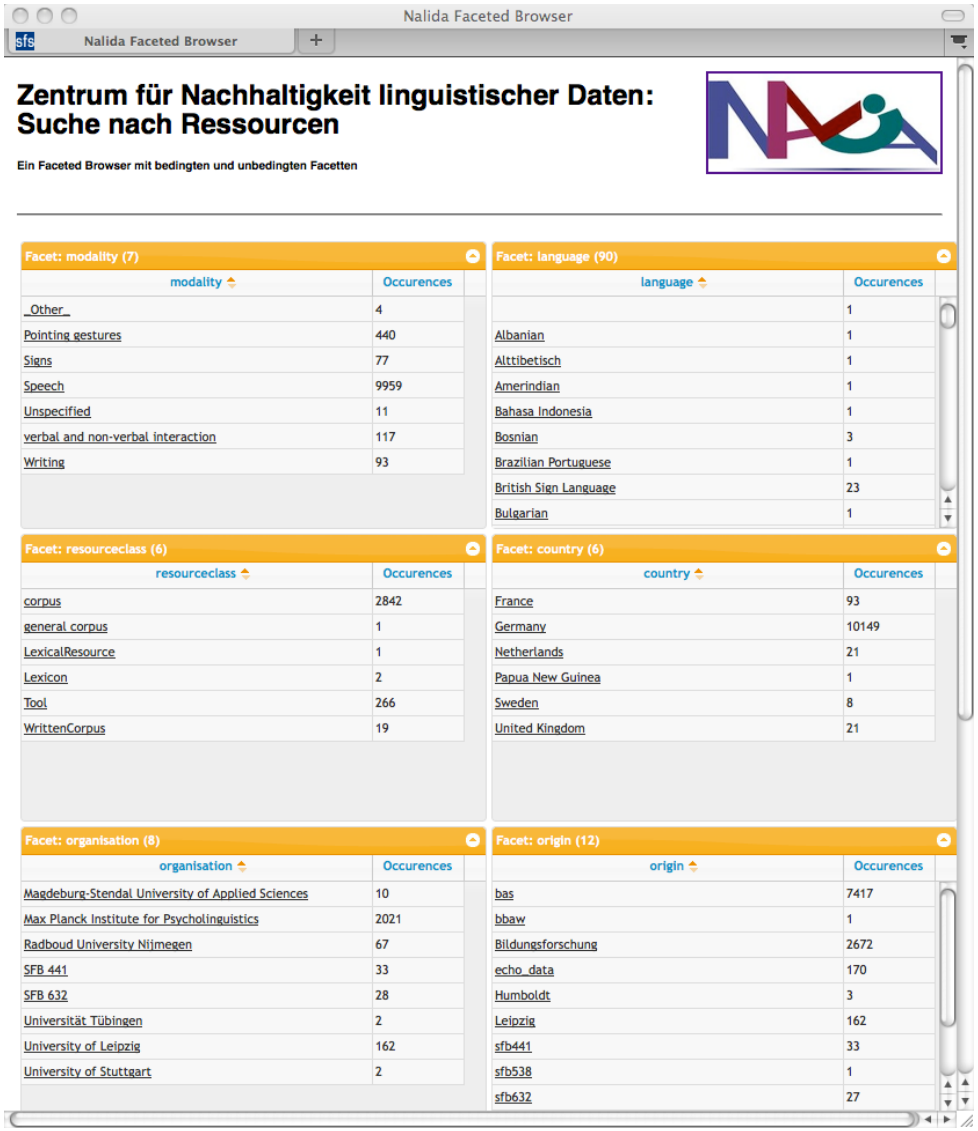
- Beschreibungsmittel gehören in die Fachwissenschaft, müssen von Wissenschaftlern verantwortet werden
- Beschreibungsmittel aus Bibliothekswissenschaft nicht ausreichend
    - Funktion der Beschreibung: Identifikation (Auffinden) und Bewertung (Nutzbarkeit zur Nachnutzung)

Metadatenframework

- *Component Metadata Infrastructure"* (CMDI, ISO 24622-1, http://www.clarin.eu/cmdi)
- **modular**
  - elementare Bausteine = standardisierte *Datenkategorien,* entnommen aus Datenkategorie-Registraturen (etwa ISOcat, siehe http://www.isocat.org, ISO 12620
  - komplexere Bausteine = Komponenten, entnommen aus der Component Registry (http://catalog.clarin.eu/ds/ComponentRegistry/)
- **flexibel**: Komponenten können an Bedürfnisse angepasst werden
  - Adaptionen werden wiederum in Component Registry verwaltet

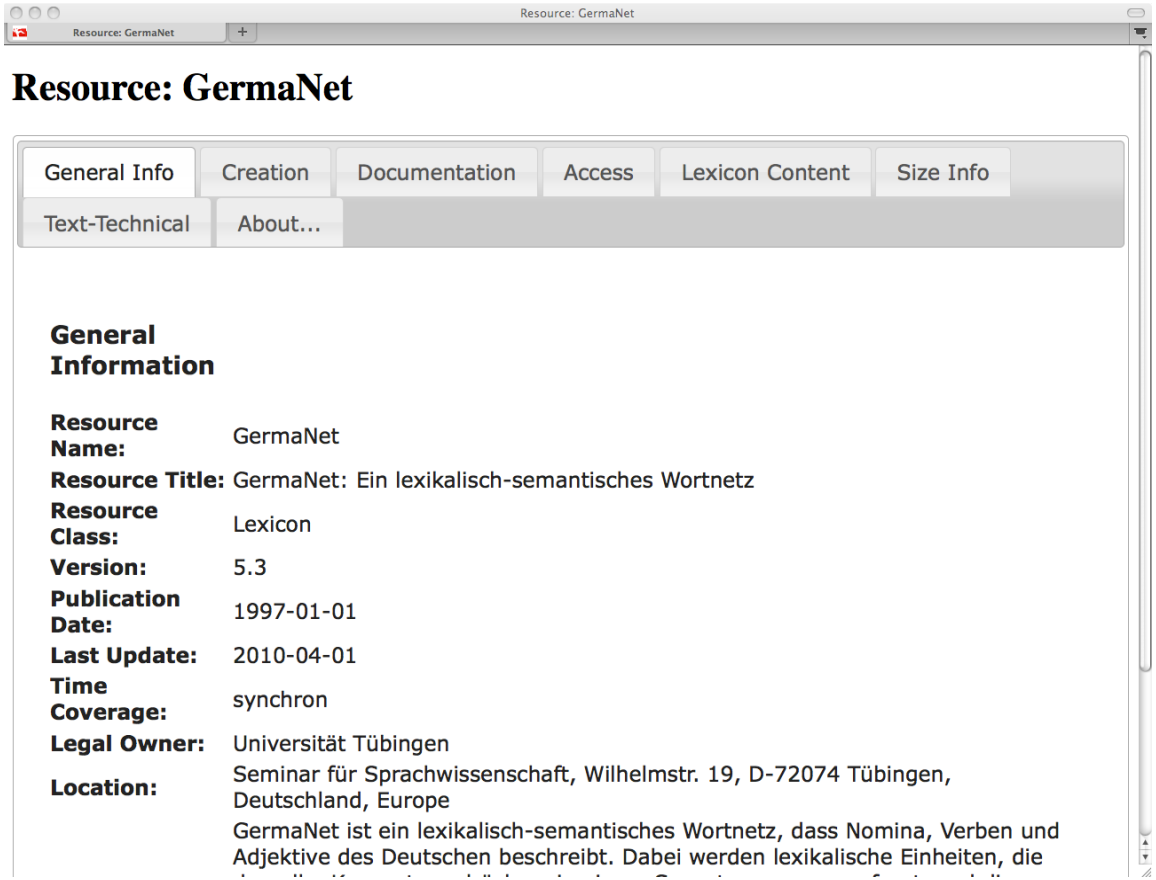
Suche Forschungsdaten

- Suche über aggregierte Datenbestände nicht trivial
- Benutzer kennt Metadatenschemata nicht
- Facetten-basierte Suche erlaubt Nutzers Exploration des Suchraums
- 6-8 Facetten, teils bedingt



Metadatensatz — Ansicht

- Abbildung zwischen Beschreibungsschema auf Karteireitern
- Zusammenfassung von Datenkategorien möglich



Link zu Faceted Browser

Archivierung in Repositorium

- Verwendung von Fedora Commons (OAIS-konform)
- nur Primärdaten mit CMDI Auszeichnung
- alle Daten über persistente Identifikatoren erreichbar
- mit OAI-PMH Schnittstelle zum “Harvesting” der Metadaten
- Derzeit ca. 50GB, 1315 Datenströme, 115 digitale Objekte

Archivierungsbedingungen

- Entscheidung für Archivierung fallbasiert
  - für die Erhebung der Forschungsdaten wurden öffentliche Fördermittel eingeworben
  - die Forschungsdaten sollen im Rahmen einer Publikation in einer wissenschaftlichen Zeitschrift zitiert werden
  - es handelt sich um eine in der Sprachwissenschaft anerkannte und oft benutzte Ressource
- Entscheidung für Archivierungsdauer ebenso kritisch

Rechtliche Fragen der Archivierung

Hinterlegungs-Vertrag

- Lizenz an Repository zur Übertragung, Speicherung, Verbreitung
- (urheber-)rechtlichen Fragen bzgl. Verbreitung und Nachnutzung
- Pflichten des Repositorium-Betreibers
    - nachhaltige Archivierung, inklusive Formatkonvertierung
    - Zugang
    - Unterhalt
    - zeitliche Vorgaben
- Personenbezogene Daten zunehmend problematisch

Verstetigung

Entwicklung einer nachhaltigen technologischen, institutionellen und personellen Infrastruktur

- Schaffung neuer Strukturen und Modelle der Zusammenarbeit von
    - **Wissenschaft**: Beschreibung von Forschungsprimärdaten mit fachwissenschaftlichen Vokabular
    - **zentralen Infrastruktur-Einrichtungen:** langfristige Vorhaltung, Verbreitung und Sicherung
- Entwicklung von **Arbeitsteilung**, **Schnittstellen,** Überprüfung auf ihre Tragfähigkeit
- Aufbau einer **prototypischen IT-Infrastruktur** für den nachhaltigen Betrieb von NaLiDa
    - Aspekte der Datensicherung (z.B. Betrieb an getrennten Standorten)
    - tragfähiges Betriebskonzept
        - Gewährung des notwendigen rechtlichen Rahmens (unter Berücksichtigung von §9 LDSG)

Verstetigung

- **NaLida als Vermittler** zwischen Fachwissenschaft und nachhaltigem Forschungsdatenmanagement
- Zusammenarbeit mit zentralen Infrastruktureinrichtungen, dem Informations-, Kommunikations- und Medienzentrum der Universität Tübingen (IKM)
    - Tübinger Universitätsbibliothek
        - stellt Handle-IDs bereit
        - macht Metadatensätze über UB-weite Suche auffindbar (EAD Untermenge, **geplant**)
            - Volltextsuche über CMDI noch nicht verfügbar, reichhaltige Information nur bei Anzeige des Metadatensatzes
    - Zentrum für Datenverarbeitung (ZDV)
        - hostet Repositorium
        - macht back-ups
        - Authentifizierung (Shibboleth)

Aufbau von Verbünden

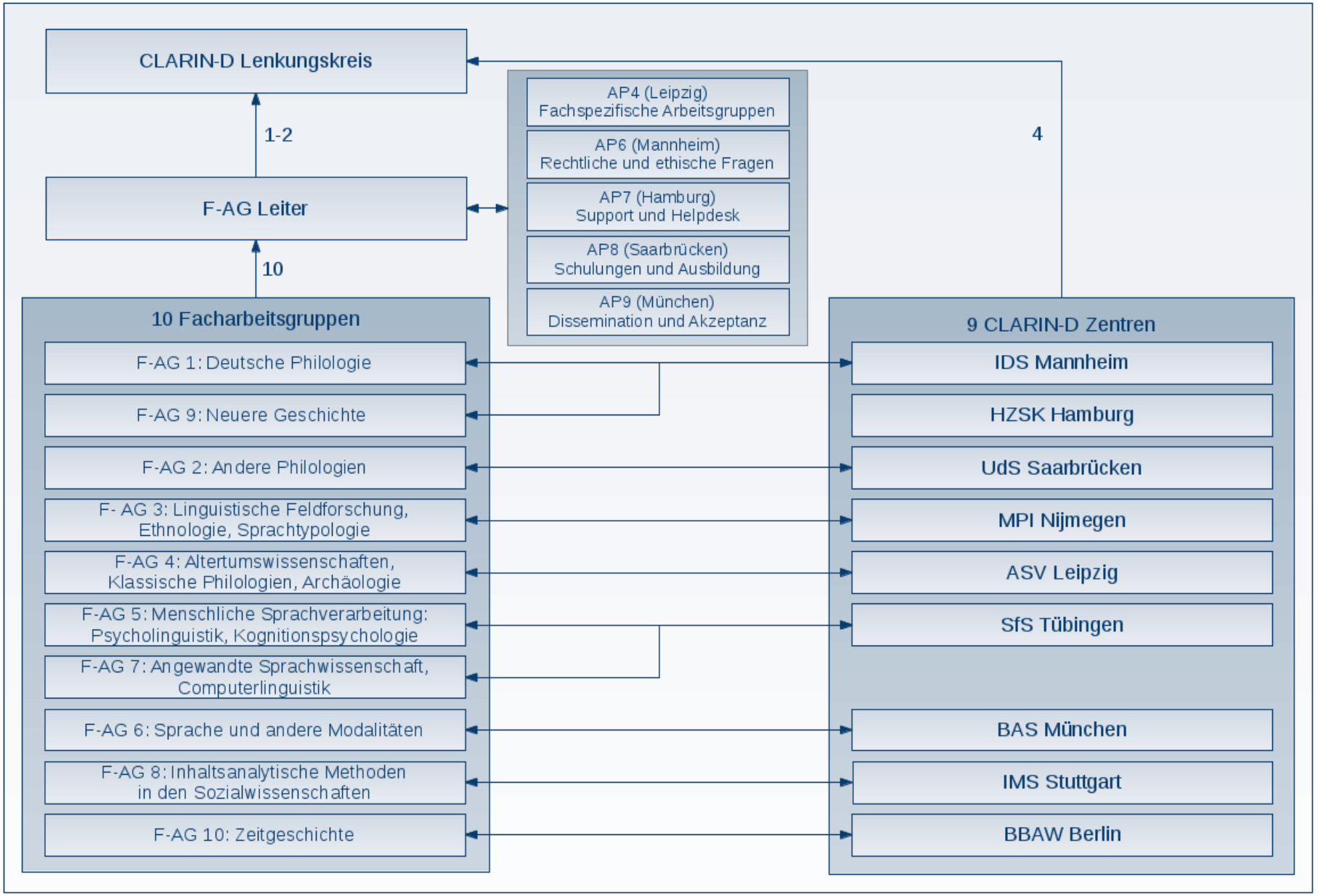
Zusammenarbeit mit **Clarin-D** und **Clarin ERIC**

- europäische Infrastruktur für Sozial- und Geisteswissenschaften
- gemeinsame Verwendung des Metadaten-Frameworks CMDI
- gegenseitiges “Harvesting” zum Metadaten-Austausch
- gemeinsame Entwicklung von Technologie & Werkzeugen
  - Metadateneditoren, Registries, VLO Faceted Browsing
  - Sprachwissenschaftliche Tools.
- personelle und institutionelle Verflechtungen hilfreich, um nationalen und europäischen Kontext in den lokalen, institutionellen Kontext einzubinden

SfS Tübingen im Clarin-D Verbund

- **Tübingen (Koord.)** – (Computerlinguistik (CL)): Annotierte Korpora, ling. Wissenskomponenten, Web-Services
- Leipzig – (Informatik): Web-Services und Textkorpora
- BBAW Berlin – (Zentrum Sprache): Deutsche Sprache: Wörterbücher, (hist.) Textkorpora
- Stuttgart – (CL): Web-Services und Textkorpora
- IDS Mannheim – (Germanistik/CL): Deutsche Sprache, Textkorpora
- LMU München – (BAS/CL) Deutsche Sprachdaten, multilinguale Daten, Web Services
- MPI Nijmegen – (Psycholinguistik): Bedrohte Sprachen, multimodale Sprachressourcen
- Hamburg – (CL): Multilinguale Sprachdaten, Transkriptionswerkzeuge
- Saarland – (CL): Multilinguale Textkorpora und Web-Services

SfS Tübingen im Clarin-D Verbund



Vertrauenswürdiges Langzeitarchiv

The Data Seal of Approval ensures that in the future, research data can still be processed in a high-quality and reliable manner, without this entailing new thresholds, regulations or high costs. The Data Seal of Approval and its quality guidelines may be of interest to research institutions, organizations that archive data and to users of that data.

Anforderungskatalog umfasst

–technische und

–formale Anforderungen

für ein nachhaltiges Datenrepositorium

- 16 Requirements, see https://assessment.datasealofapproval.org/media/files/DSA_booklets/DSA-booklet_1_June2010_1.pdf

CLARIN CENTRE B

Data Seal of Approval

The *data producer*

1. deposits the research data in a data repository with sufficient information for others to assess the scientific and scholarly quality of the research data and compliance with disciplinary and ethical norms

2. provides the research data in formats recommended by the data repository

3. provides the research data together with the metadata requested by the data repository

Data Seal of Approval

The *data repository*

4. *has an explicit mission in the area of digital archiving and promulgates it*
5. uses due diligence to ensure compliance with legal regulations and contracts
6. applies documented processes and procedures for managing data storage
7. has a plan for long-term preservation of its digital assets
8. Archiving takes place according to explicit workflows across the data life cycle
9. assumes responsibility from the data producers for access to and availability of the digital objects
10. enables the users to utilize the research data and refer to them
11. ensures the integrity of the digital objects and the metadata
12. ensures the authenticity of the digital objects and the metadata

Data Seal of Approval

The technical infrastructure

13. explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS

The *data consumer*

14. must comply with access regulations set by the data repository

15. conforms to and agrees with any codes of conduct that are generally accepted in higher education and research for the exchange and proper use of knowledge and information

16. respects the applicable licences of the data repository regarding the use of the research data

Vielen Dank

für ihre Aufmerksamkeit