Datenübernahme am Text+ Datenzentrum des Leibniz-Instituts für Deutsche Sprache, Mannheim

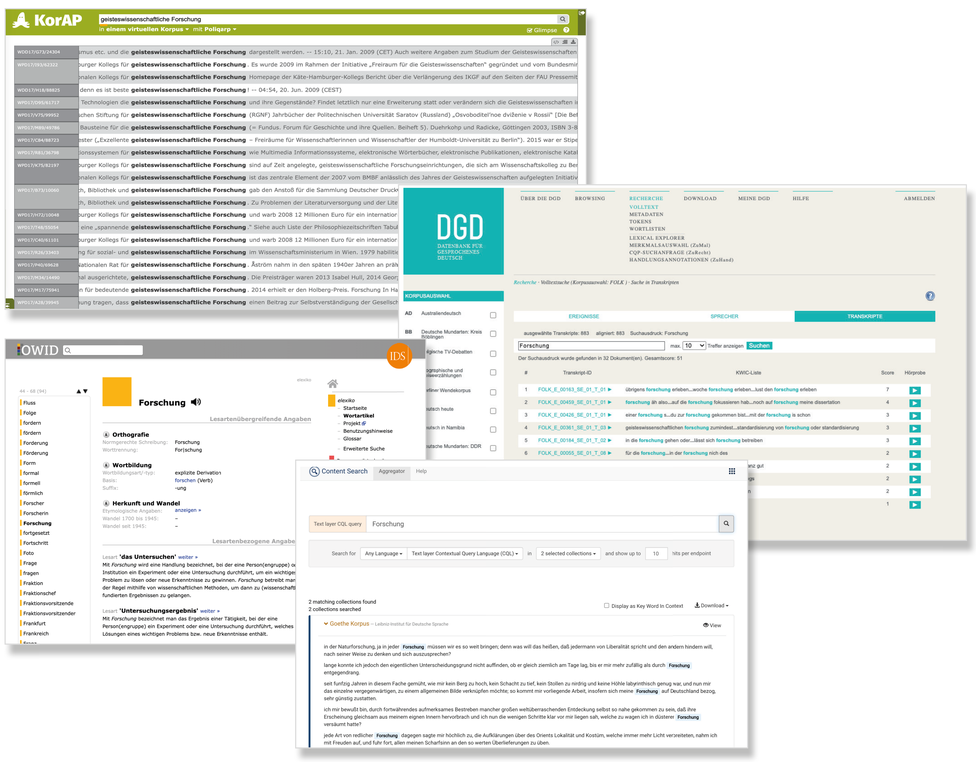
Andreas Witt, Antonina Werthmann, Thorsten Trippel

Beispiele vorhandener Daten am IDS

- Deutsches Referenzkorpus – DeReKo
  - 55 Milliarden Wörter (Stand März 2023)
  - Größte linguistisch motivierte Sammlung elektronischer Korpora
  - Geschriebene deutschsprachige Texte aus Gegenwart und neuerer Vergangenheit
  - Verschiedene Textarten: belletristische, wissenschaftliche und populärwissenschaftliche Texte, Zeitungstexte, ...
- Archiv für Gesprochenes Deutsch (AGD)
  - Datenbank für Gesprochenes Deutsch (DGD)
  - Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK)
- Lexikalische Daten: Online-Wortschatz-Informationssystem Deutsch (OWID)
- Weitere Ressourcen (insbesondere zur Langzeitarchivierung)
  - BOLSA (Bonner Längsschnittstudie des Alterns)
  - DISKO (Deutsch im Studium: Lernerkorpus): schriftliches Lernerkorpus
  - MIKO (Mitschreiben in Vorlesungen: Ein multimodales Lehr-Lernkorpus): multimodales, wissenschaftssprachliches Vorlesungskorpus; mit Audio- und Videoaufnahmen sowie geschriebener Sprache

Text+ Datenzentrum am Leibniz-Institut für Deutsche Sprache, Mannheim	
Spezialisierung	▸ Neuhochdeutsch • gesprochene Sprache • geschriebene Sprache ▸ linguistisch annotierte Korpora ▸ Deutsch in nicht-primär deutschsprachigen Ländern ▸ ...
Modalität	▸ geschrieben ▸ gesprochen
Akzeptierte Datenformate	▸ Textuelle Datenformate: • TEI (I5 Format, siehe https://www.ids-mannheim.de/digspra/kl/projekte/korpora/textmodell) ▸ Gesprochensprachliche Datenformate: • Transkription gesprochener Sprache nach ISO/TEI (ISO 24624:2016) • Signaldateien in den üblichen Formaten (WAV, ...) ▸ Weitere Formate nach Absprache
Ansprechpersonen	▸ Andreas Witt ▸ Thorsten Trippel ▸ Antonina Werthmann

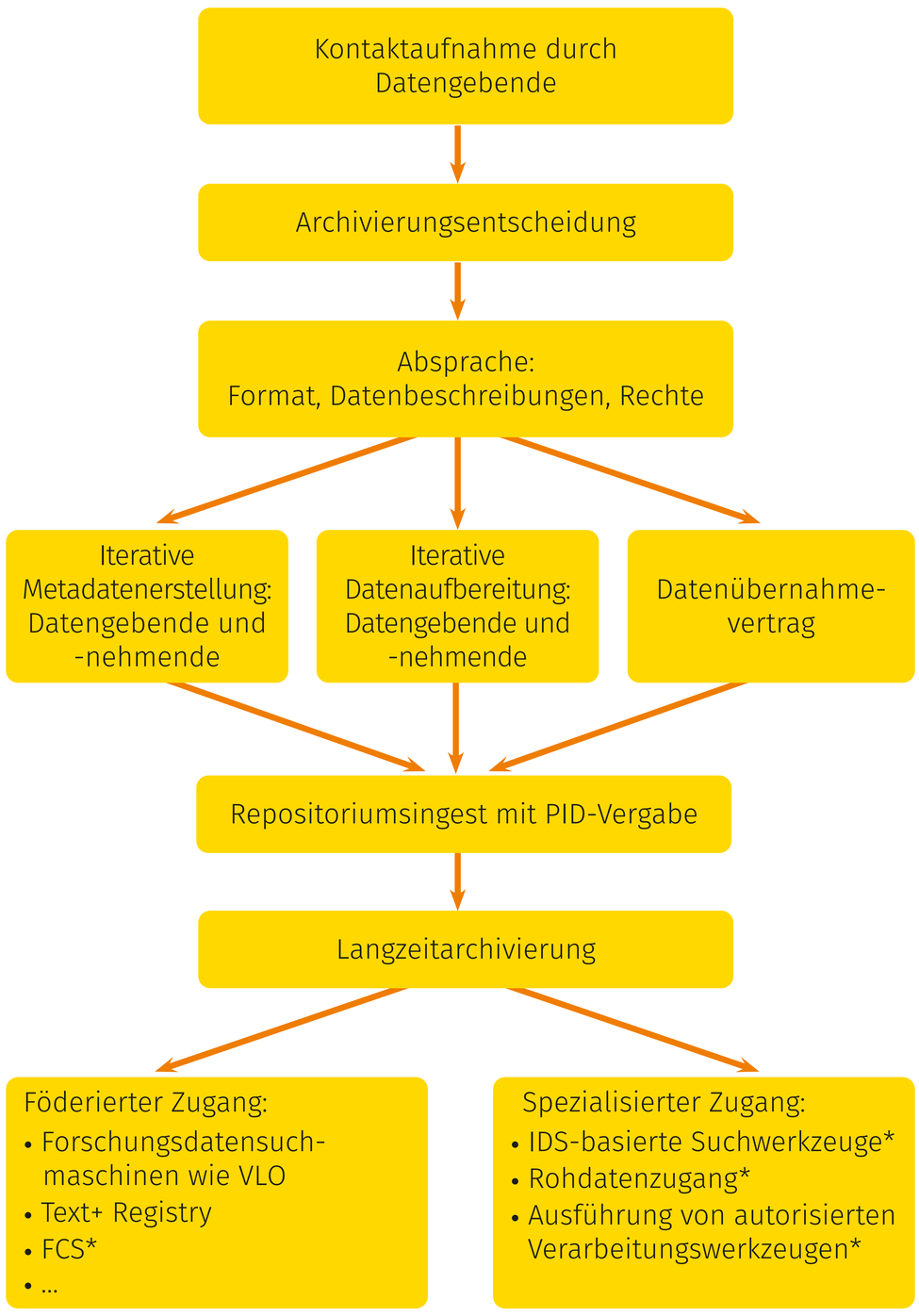
Vorhandene (spezialisierte) Datensuchfunktionen



Daten Einbringen

Voraussetzung:

- Unterstützte Datentypen
  - Qualitätsgesichert
  - Auswahlprozess
  - Geklärte Rechte
- Datenüberlassungsvertrag
- Offene Lizenz: CC-BY 4.0 oder höher
- Metadaten nach ISO 24622 1/2
- Detaillierte Datenübernahme-richtlinien siehe DOI: 10.14618/ids-pub-8791
- Abweichungen nach Absprache möglich



* In Abhängigkeit von rechtlichen Möglichkeiten/Lizenzen, Datentypen, Formaten, Umfängen

FAIR

F wie Findable

- Auffindbar mit persistentem Identifikator (PID)
  - Aktuell: Handle
  - Zukunft: DOI
- Nachweis über Forschungsdatensuchmaschinen, z. B. VLO, und Auffindbarkeit über Standardsuchmaschinen

A wie Accessible

- Downloadmöglichkeit freier Daten
- Zugang zu vorhandenen zugangsbeschränkten Daten
- Metadaten frei zugänglich über technische Protokolle, auch über die Verfügbarkeit der Daten hinaus

I wie Interoperable

- Je nach Datentyp
- Nutzbar in den vorhandenen Auswertungswerkzeugen
- (Meta-)daten: standardkonform und maschinell interpretierbar

R wie Reusable

- Nutzbar unter den Bedingungen der Lizenzen in eigenen Werkzeugen
- Verwendung in anderen Projektkontexten
- Klare Rechte an den Daten
- Vertrauenswürdige Aufbewahrung
- Langzeitarchivierung

Kontakt:
Prof. Dr. Andreas Witt /
Dr. Antonina Werthmann /
Dr. Thorsten Trippel
Abteilung Digitale Sprachwissenschaft
Leibniz-Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim
{trippel|werthmann|witt}
@ids-mannheim.de

Hausanschrift:
Leibniz-Institut für Deutsche Sprache
R 5, 6-13
68161 Mannheim

Tel.: +49 621 1581-0
Fax: +49 621 1581-200
info@ids-mannheim.de
www.ids-mannheim.de

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-121116