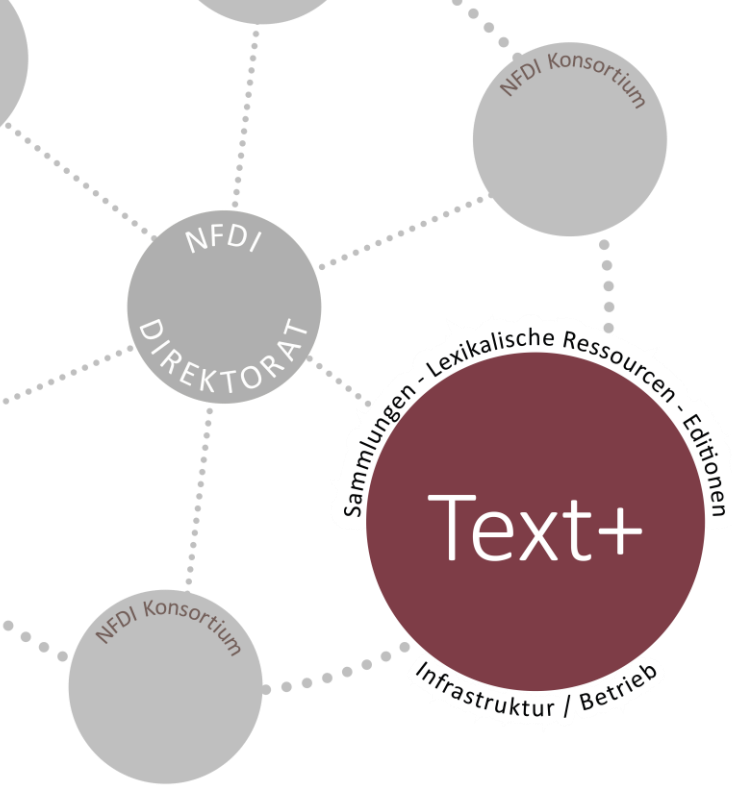
Präsentiert durch [Name(n) der Institution(en)/Projekt(e)] im Rahmen von Text+. Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 460033370



Wohin damit?
Storing and reusing my language data

Minute Madness der Datenzentren

text-plus.org
office[at]text-plus.org

1. Universität zu Köln
Data Center for the Humanities

Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

- Spezialisierung in Text+:
  - Audiovisuelle Sprachdaten
  - Außereuropäische Sprachen
  - Sprachdokumentation
  - Oral Literature
  - Minderheitensprachen
- allg. FDM Support, FDM Beratung, Archivierungsservices
- https://dch.phil-fak.uni-koeln.de/fdm-services

LAC Language Archive Cologne

CORE TRUST SEAL

CLARIN K CENTRE

CLARIN Knowledge Centre
CKLD

Text+

2. SUB Göttingen

TextGrid Repository

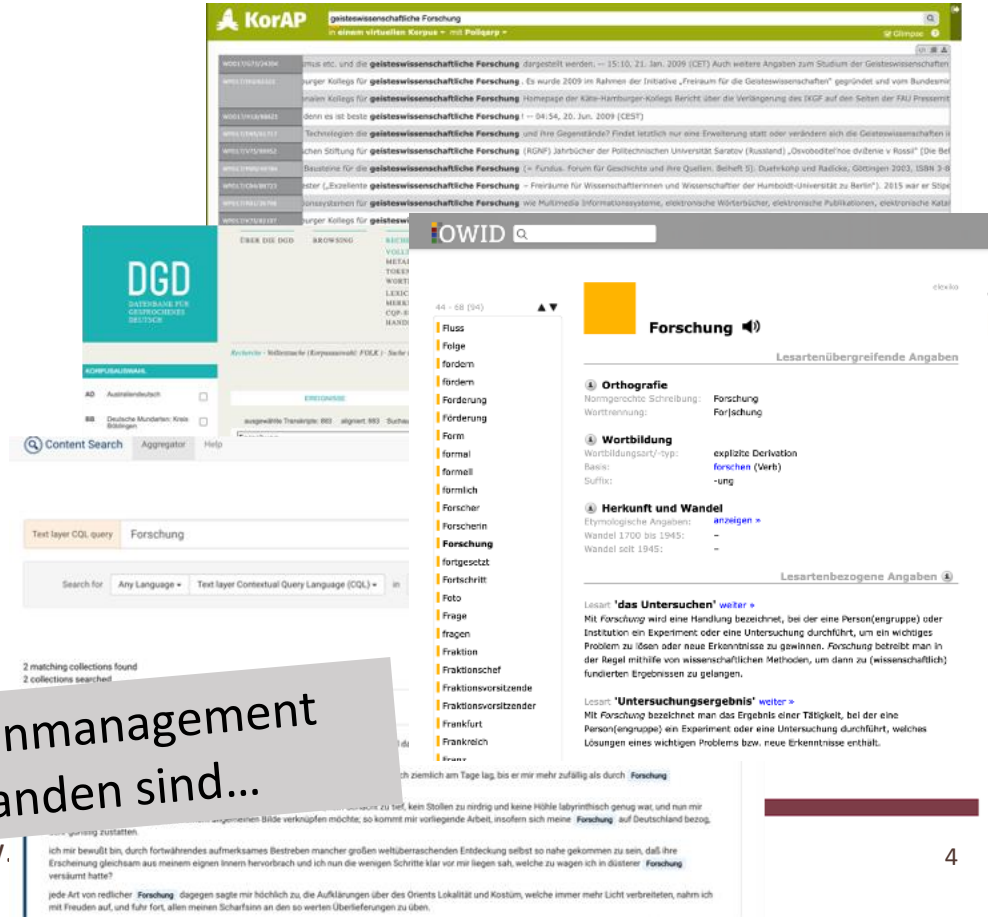
DARIAH-DE

- **TextGrid Repository** seit 2011
- Fokus auf XML/TEI-kodierter Texte und Bilder
- Ideal geeignet für digitale Editionen

- **DARIAH-DE Repository** seit 2017
- Dauerhaft gültige Persistente Identifikatoren (DOI) für alle gespeicherten Objekte
- Nachhaltige und sichere Archivierung von Datensammlungen (Kollektionen).

3. Leibniz-Institut für Deutsche Sprache

- Neuhochdeutsch
  - gesprochen
  - geschrieben
- Korpora
  - Deutsches Referenz Korpus (DeReKo)
  - Archiv Gesprochenes Deutsch (AGD)
  - …
- Lexikalische Ressourcen
  - OWID
- Deutsch in nicht-primär deutschsprachigen Ländern



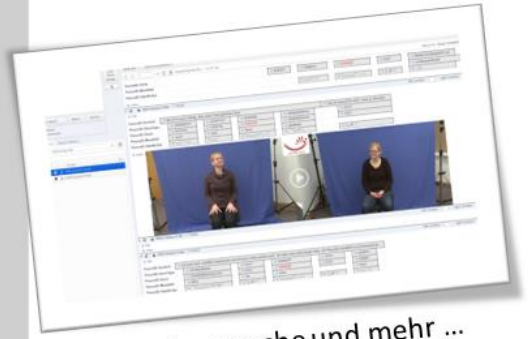
Wir können auch Forschungsdatenmanagement für Daten, die nicht am IDS entstanden sind…

www.

Text+ Sammlungen, Lexikalische Ressourcen, Editionen – Infrastruktur / Betrieb

4.

Gesprochene Sprache und mehr …



Gebärdensprache und mehr …

UHH

(HZSK)

AKADEMIE DER WISSENSCHAFTEN IN HAMBURG

Zentrum für nachhaltiges Forschungsdatenmanagement der UHH

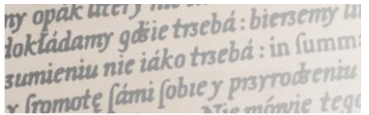
5. Universität des Saarlandes

» **CLARIND-UdS:** Annotierte Korpora, geschriebener Sprache, insbesondere fremdsprachliche und multilinguale Sprachkorpora (parallele und vergleichbare Korpora) und registerspezifische Korpora.

» Datenformate: vrt (Open Corpus Workbench, CQPweb)

Lynch . The first acquaintance I had with him was at Bath	last winter	was 12 months . I was frequently in his company , and
his own , and had the charge of sixteen or seventeen ships	every winter	; and has the character of a very honest man . Acquitted
What are you ? Prisoner . I am a bricklayer , and	in winter	time I follow chair work . I live by Grosvenor-square . Q.
to the coal trade , but it is customary for him in	the winter	time to go to Rotterdam . Q. Did you ever hear of
have seen him at Horsey before then , both in the summer	and winter	at my house , there I am a farmer . I have
for such Practices , and her mother was burnt for the same	the	have_VH0 seen_VVN him_PPHO1 at_II Horsey_JJ before_II then_RT ,_, both_RR in_II the_AT summer_NNT1 and_CC winter_NNT1 at_II my_APPGE house_NN1 ,_, there_RL I_PPIS1 am_VBM a_AT1 farmer_NN1 ._. I_PPIS1 have_VH0 cast , and afterwards transported
so that he could not lie down in his bed all	the	s ; he turned yellow , and
of November . I live in May-fair-market , am a Chairman in	the winter	; and laboured in summer . I was at work upon a
not tell which month Christmas was in , but knows it is	in winter	time . Elizabeth Headland , the mother to the last evidence ,
Christmas-day , or what day of the week , or whether it	was winter	or summer , but yet she went to church on new Christmas-day
I went and brought her the things and change . In	the winter	her mother was lame ; she told me her daughter had found

GECCO
GERMAN ENGLISH CONTRASTS IN COHESION



CLARE
CORPUS LATINUM REFERENS

6. Deutsches Textarchiv an der BBAW

Historische Textsammlungen im Deutschen Textarchiv (DTA) an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW)

1600–1920

Link www.deutschestextarchiv.de | Link www.dwds.de/d/korpora/dtaxl

Sprache deu | Format DTABf | Format TCF

Sammlungen > 40 | Transkription double keying, ocr

Genre Belletristik/Gebrauchsliteratur/Wissenschaft/Zeitung | Lizenz CC BY-SA 4.0

Das DTA ist ein aktives Archiv für deutschsprachige, historische Korpora und Sammlungen am Zentrum Sprache der BBAW. Es umfasst annotierte Volltexttranskriptionen von Drucken, Zeitungen und Zeitschriften sowie handgeschriebene Dokumente verschiedener Gattungen und Textarten. Für die Transkriptionen bietet das DTA mit dem DTABf – ein Subset der TEI-P5-Guidelines – ein etabliertes Basisformat an.

https://github.com/deutschestextarchiv/

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

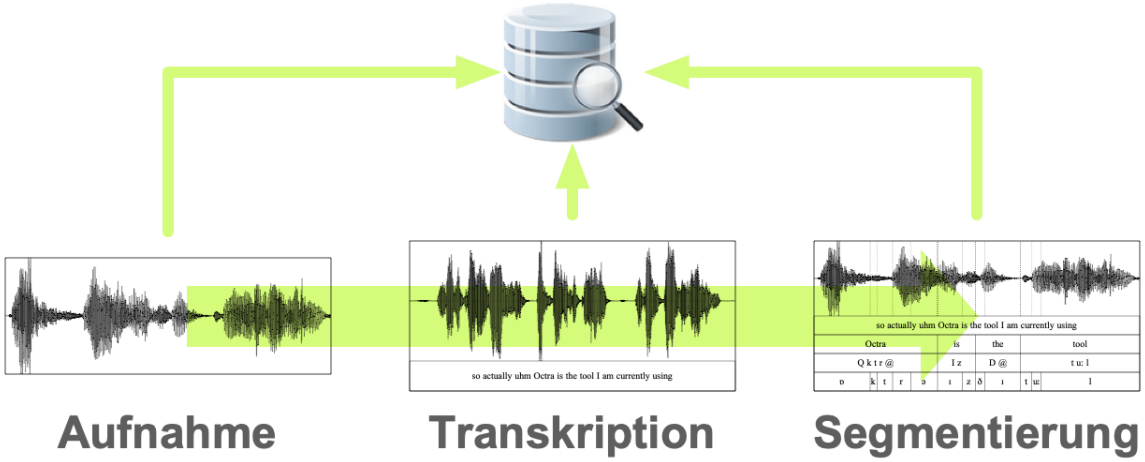
D T A

D W D S

7. LMU München, Bayerisches Archiv für Sprachsignale

» CTS-zertifiziertes Repository mit 50+ annotierten Sprachkorpora

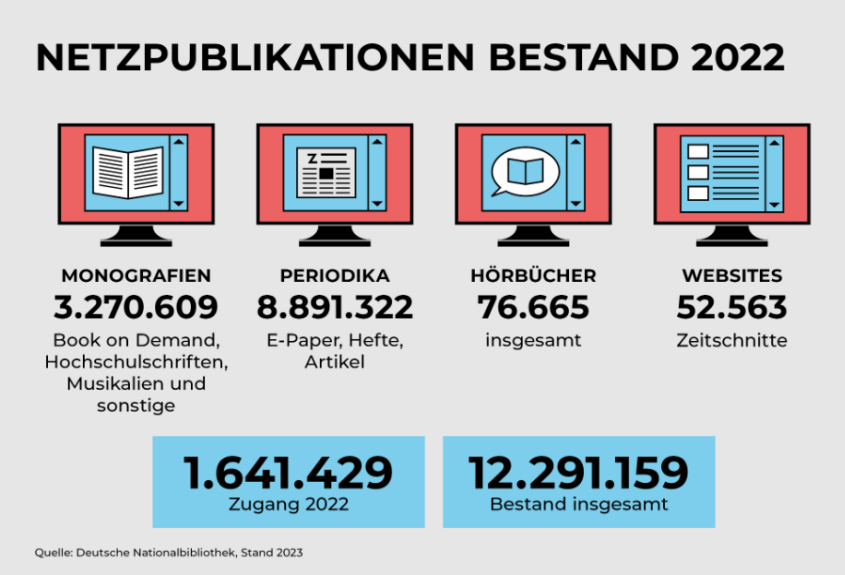
» WebDienste für sprachtechnologische Aufgaben



IFCASL, LangGener & Whisper KI

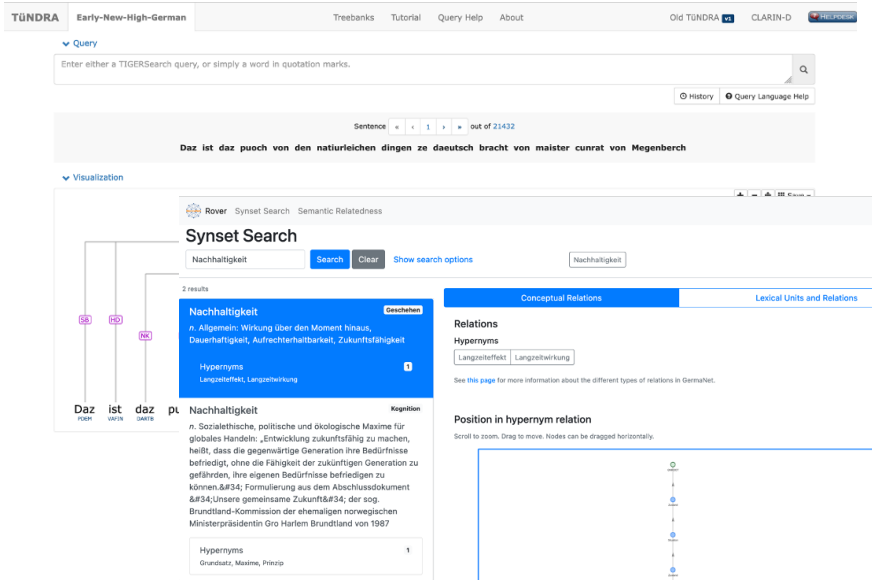
8. Deutsche Nationalbibliothek

- Alle seit 1913 in Deutschland erschienenen Publikationen und Tonträger sowie E-Books, E-Paper, Online-Hochschulschriften, Noten, Websites etc.
- DH-Call: Unterstützung beim Zugang zu urheberrechtlich geschütztem Material
- Heute über 12 Mio. Netzpublikationen im Bestand
- Analyse und Vernetzung von über 200 Mio. Metadaten aus Deutschland und Österreich z.B. mit Culturegraph
- Normdaten der Gemeinsamen Normdatei (GND)



9. Universität Tübingen

- Forschungsdatenmanagement für
  - SFB 833, 441, GRK 1808 (ausgelaufen)
  - Seminar für Sprachwissenschaft
  - Daten von anderen Datengebenden
- Spezialisiert auf
  - Baumbanken/Treebanks
  - Wortnetze
  - Word-Embeddings
  - geschrieben
  - gesprochen

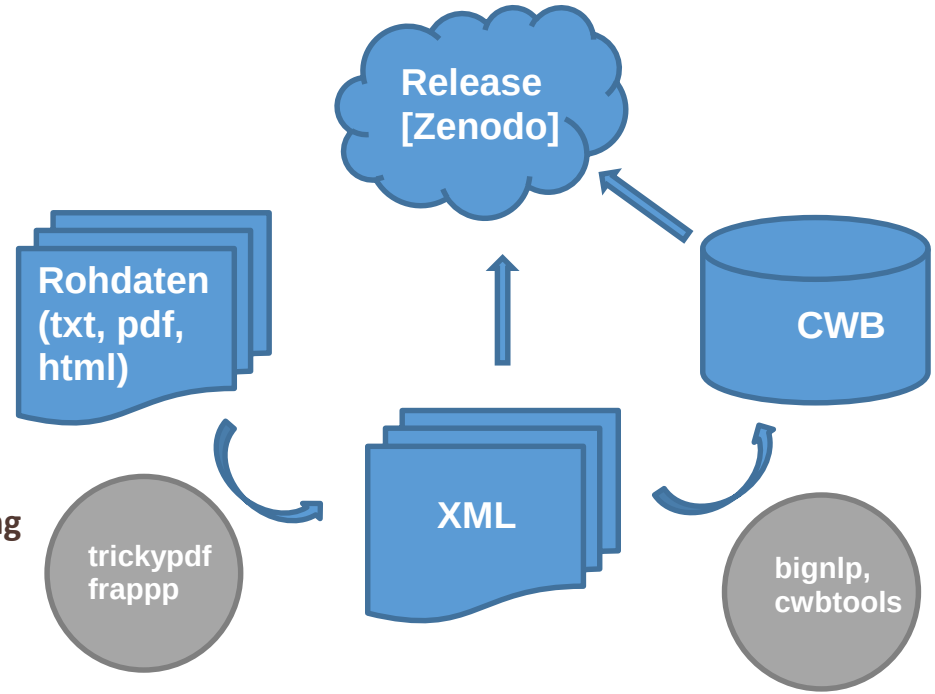


Zertifikat erteilt am
01. Juni 2023

Rezertifizierung CLARIN-B
im Verfahren

10. Universität Duisburg-Essen

- **Kontext “PolMine Project” (polmine.de)**
  Data & Code for Corpus Analysis
- **Gemeinfreie Daten als Schwerpunkt**
  Spezialisierung auf Parlamentsprotokolle
- **GermaParl v2.0.0 (Release 05/2023)**
  alle Bundestagsdebatten 1949-2021
- ***Open Source*-Toolset für die Korpusaufbereitung**
  R-Pakete verfügbar über GitHub / CRAN
- **Prozessberatung als Kompetenzzentrum**
  Workflows und Best Practices für reproduzierbare Aufbereitung gemeinfreier Textdaten



11. Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

- Dieses **Text+ Langzeitarchiv** wird aktuell aufgebaut und soll 2023 an den Start gehen.
- Es bietet eine Infrastruktur für den langfristigen und sicheren Erhalt von Daten mit geisteswissenschaftlichem Bezug.
- Es wird *bitstreampreservation* bieten und das unkomplizierte Archivieren von Daten ermöglichen, ohne Einschränkung des Datenformats.
- Sowohl offene als auch geschützte Archivbereiche werden möglich sein.