

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>



CLARIN Knowledge Centre



## 2. SUB Göttingen

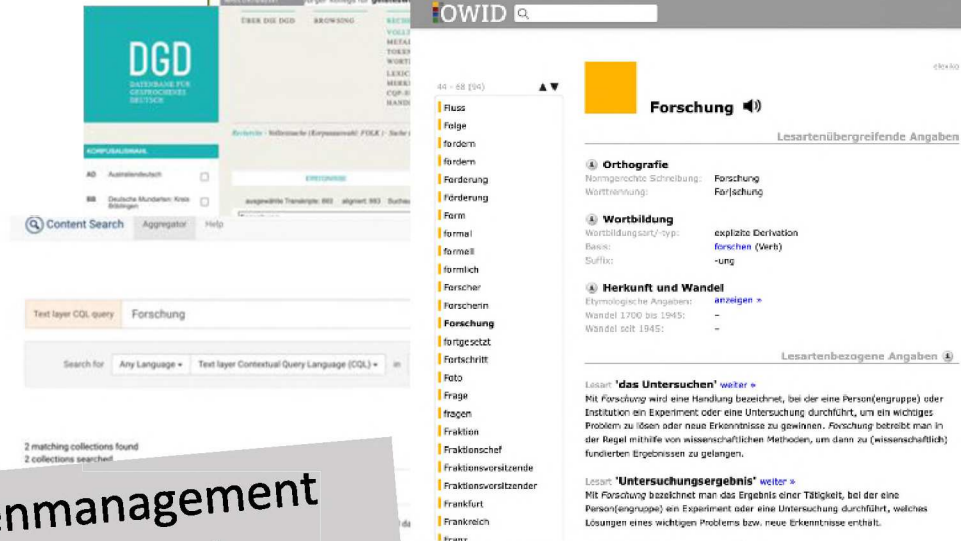


- **TextGrid Repository** seit 2011
- Fokus auf XML/TEI-kodierter Texte und Bilder
- Ideal geeignet für digitale Editionen

- **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



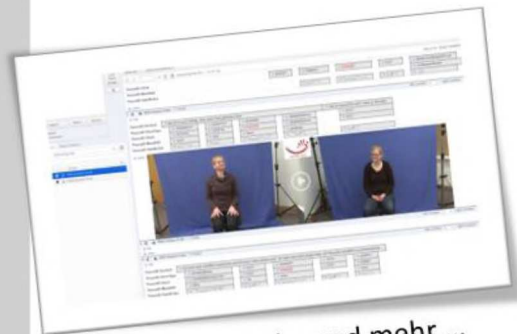
**Wir können auch Forschungsdatenmanagement für Daten, die nicht am IDS entstanden sind...**

www.

4.



Gesprochene Sprache und mehr ...



Gebärdensprache und mehr ...



Zentrum für nachhaltiges  
Forschungsdatenmanagement der UHH

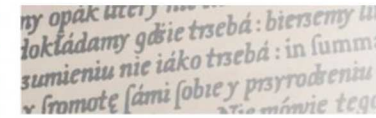


# 5. Universität des Saarlandes

- » **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	<b>last winter</b>	was 12 months . I was frequently in his company , and
his own , and had the charge of sixteen or seventeen ships	<b>every winter</b>	; and has the character of a very honest man . Acquitted
What are you ? Prisoner . I am a bricklayer , and	<b>in winter</b>	time I follow chair work . I live by Grosvenor-square . Q.
to the coal trade , but it is customary for him in	<b>the winter</b>	time to go to Rotterdam . Q. Did you ever hear of
have seen him at Horsey before then , both in the summer	<b>and winter</b>	at my house , there I am a farmer . I have
for such Practices , and her mother was burnt for the same	<b>the</b>	have_VHO seen_VVN him_PPHO1 at_II Horsey_JJ
so that he could not lie down in his bed all	<b>the</b>	before_II then_RT_ , both_RR in_II the_AT
of November . I live in May-fair-market , am a Chairman in	<b>the winter</b>	summer_NNT1 and_CC winter_NNT1 at_II my_APPGE
not tell which month Christmas was in , but knows it is	<b>in winter</b>	house_NN1 _ , there_RL_I_PPIS1 am_VBM a_AT1
Christmas-day , or what day of the week , or whether it	<b>was winter</b>	farmer_NN1 _ , I_PPIS1 have_VHO
I went and brought her the things and change . In	<b>the winter</b>	and labourer in summer . I was at work upon a
	<b>in winter</b>	time . Elizabeth Headland , the mother to the last evidence ,
	<b>was winter</b>	or summer , but yet she went to church on new Christmas-day
	<b>the winter</b>	her mother was lame ; she told me her daughter had found



**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](http://www.deutschestextarchiv.de) [Link www.dwds.de/d/korpora/dtaxl](http://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**

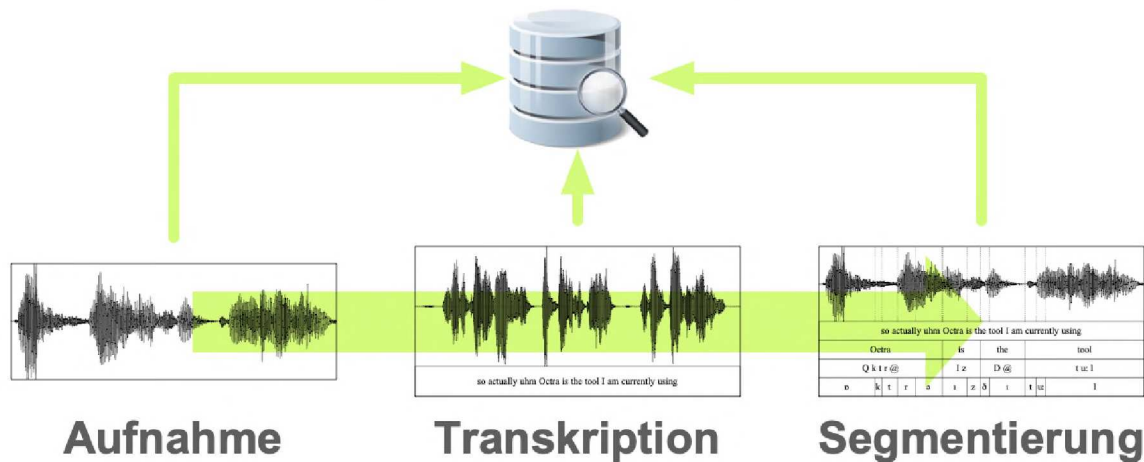
Das DTA ist ein aktives Archiv für deutschsprachige, historische Korpora und Sammlungen am Zentrum Sprache der BBAW. Es umfasst annotierte Volltexttranskriptionen von Drucken, Zeitungen und Zeitschriften sowie handgeschriebene Dokumente verschiedener Gattungen und Textarten. Für die Transkriptionen bietet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

<https://github.com/deutschestextarchiv/>



# 7. LMU München, Bayerisches Archiv für Sprachsignale

- » CTS-zertifiziertes Repository mit 50+ annotierten Sprachkorpora
- » WebDienste für sprachtechnologische Aufgaben

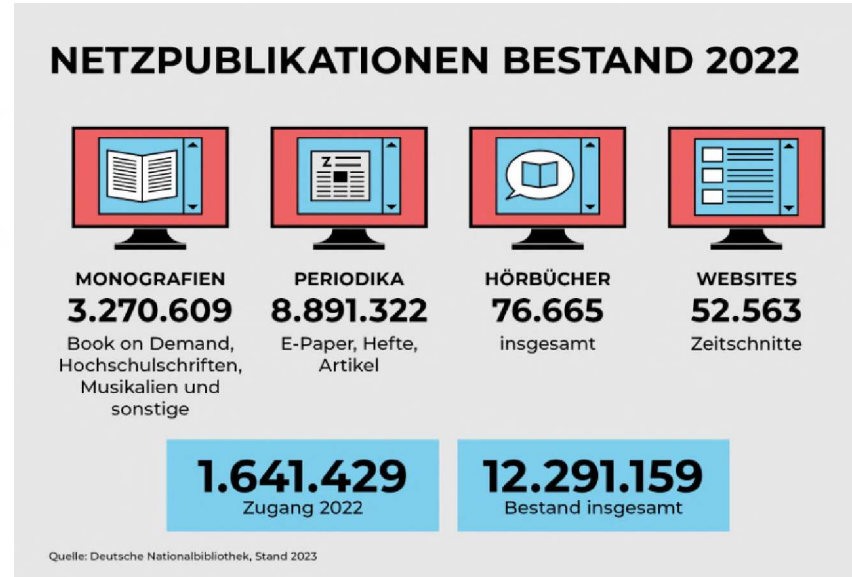


**IFCASL,  
LangGener &  
Whisper KI**



# 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 Datengebern
- » Spezialisiert auf
  - » Baumbanken/Treebanks
  - » Wortnetze
  - » Word-Embeddings
  - » geschrieben
  - » gesprochen

The screenshot shows the TUNDRA search interface. At the top, there are navigation links for 'TUNDRA', 'Early-New-High-German', 'Treebanks', 'Tutorial', 'Query Help', and 'About'. Below this is a search bar with the text 'Enter either a TIGERsearch query, or simply a word in quotation marks.' and a search button. The search results are displayed in a table with columns for 'Sentence', 'Visualization', and 'Results'. The first result is a sentence: 'Daz ist daz puoch von den natuerlichen dingen ze daeutech bracht von maister curnat von Megeberch'. Below this, there is a visualization of the search results, showing a tree structure with nodes for 'Daz', 'ist', 'daz', and 'pt'. The main content area shows the search results for 'Nachhaltigkeit', including a list of related terms and a detailed description of the concept.

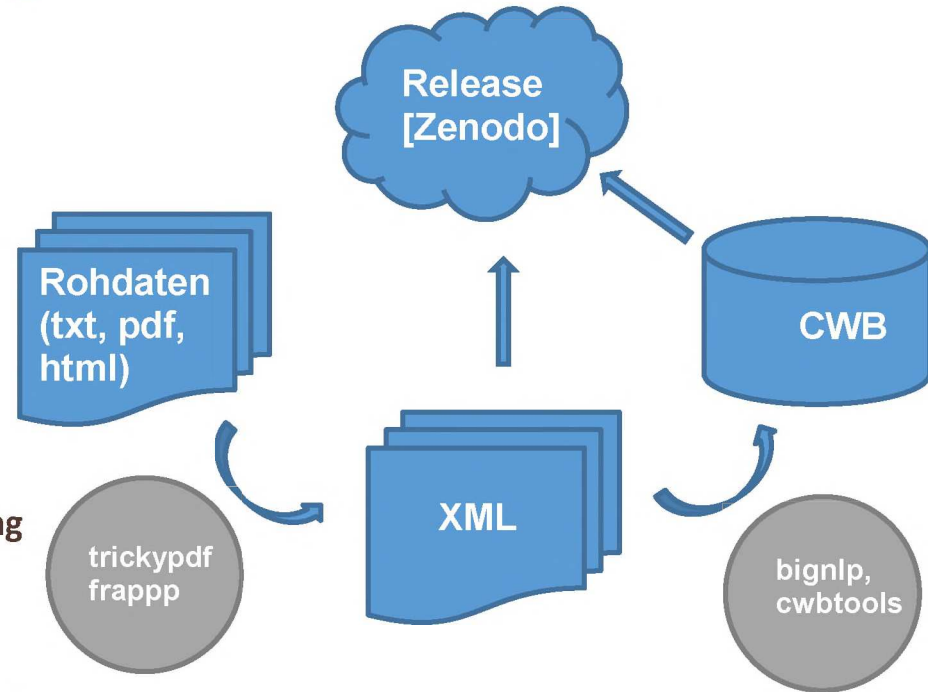
Zertifikat erteilt am  
01. Juni 2023

Rezertifizierung CLARIN-B  
im Verfahren



# 10. Universität Duisburg-Essen

- » **Kontext “PolMine Project”** ([polmine.de](http://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.