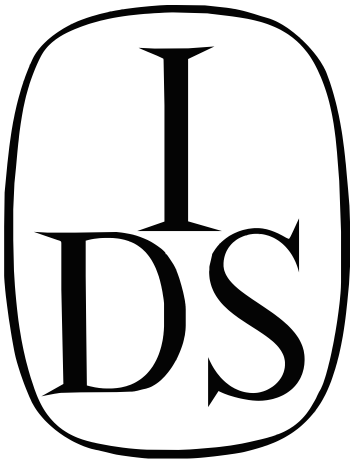
Forschungsdaten in den Geisteswissenschaften – die germanistische Sprachwissenschaft

Andreas Witt

Institut für Deutsche Sprache, Mannheim

Workshop Forschungsdaten
WGL Geschäftsstelle Berlin
2012-05-10

Institut für Deutsche Sprache



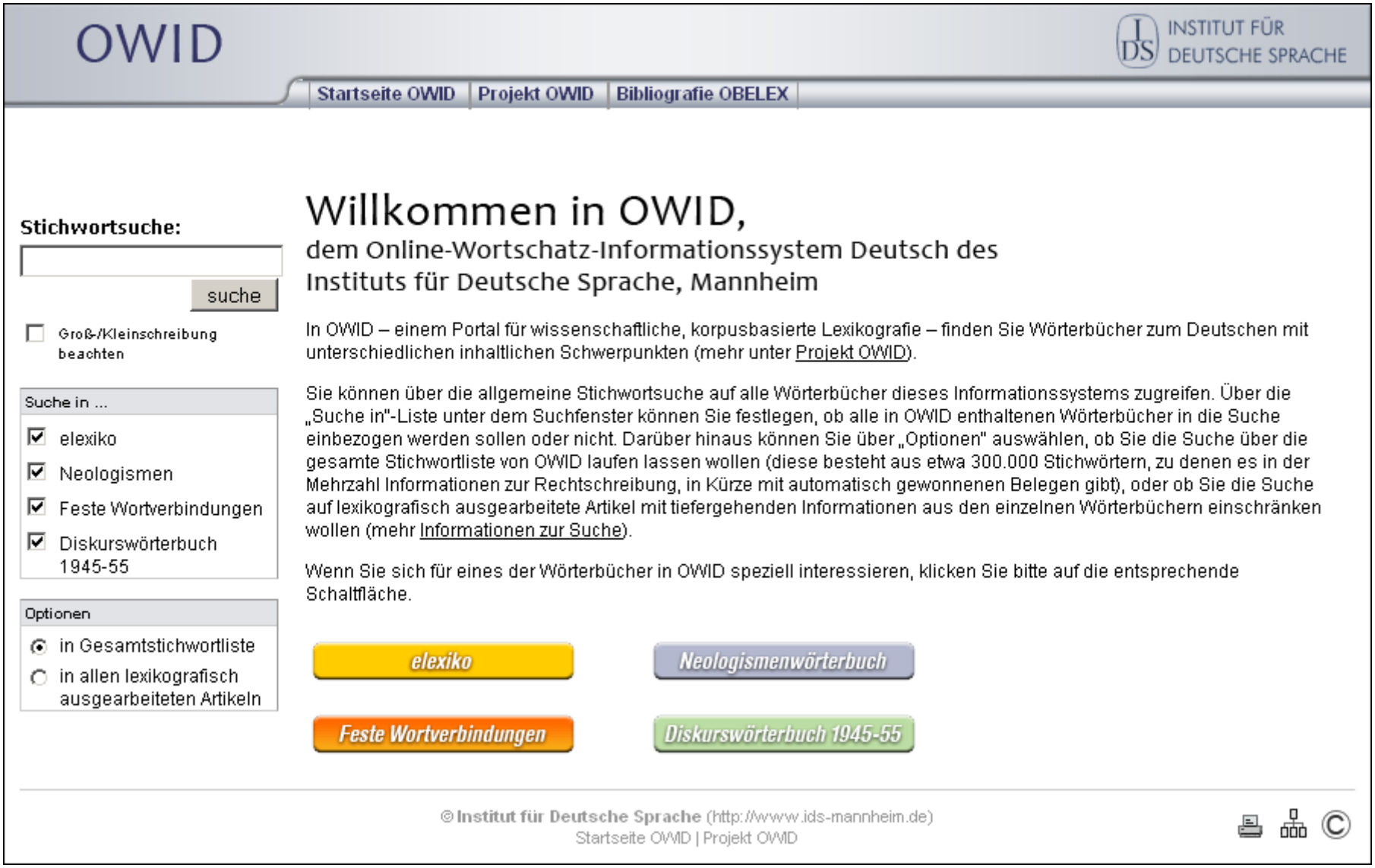
- gegründet 1964
- primäres außeruniversitäres Forschungsinstitut zur Erforschung und Dokumentation der deutschen Gegenwartssprache

Linguistische Primärdaten am IDS

- Lexikographische Informationssysteme
  - z.B. OWID
- Geschriebensprachliche Korpora
  - z.B. DeReKo
- Gesprochensprachliche Korpora
  - z.B. FOLK

Online-Wortschatz-Informationssystem Deutsch (OWID)

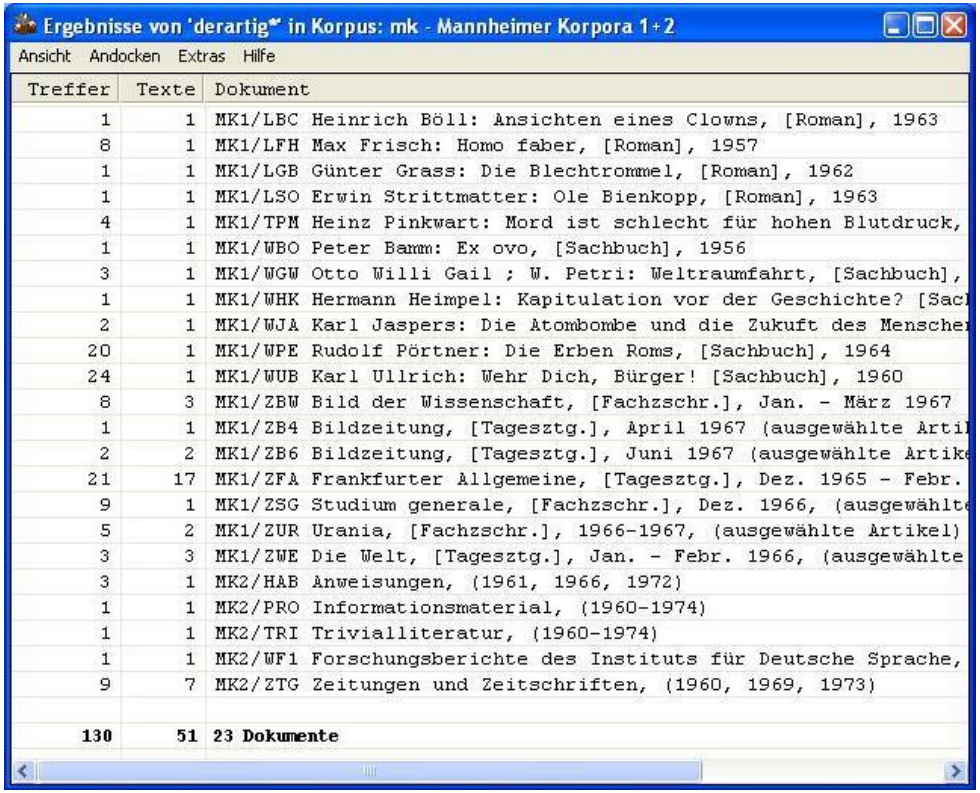
- Plattform für ein lexikografisches Informationssystem
- Bereitstellung lexikografischer und lexikologischer Arbeiten in computerlexikografisch angemessener Weise
- Erforschung der gemeinsamen Präsentierbarkeit unterschiedlicher lexikografischer Ressourcen, der Vernetzbarkeit von Ressourcen und des Umgangs der Benutzer mit neuen Formen eines Wortschatzinformationssystems

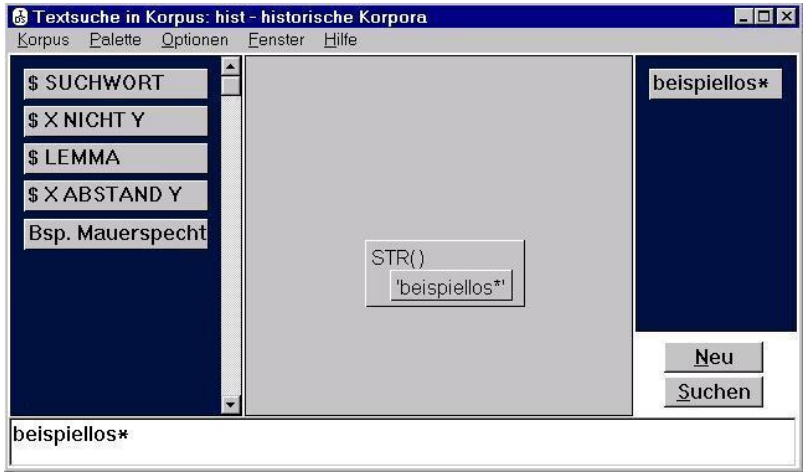

OWID
INSTITUT FÜR DEUTSCHE SPRACHE
Startseite OWID
Projekt OWID
Bibliografie OBELEX
Stichwortsuche:
suche
Groß-/Kleinschreibung beachten
Suche in ...
elexiko
Neologismen
Feste Wortverbindungen
Diskurswörterbuch 1945-55
Optionen
in Gesamtstichwortliste
in allen lexikografisch ausgearbeiteten Artikeln
Willkommen in OWID,
dem Online-Wortschatz-Informationssystem Deutsch des Instituts für Deutsche Sprache, Mannheim
In OWID – einem Portal für wissenschaftliche, korpusbasierte Lexikografie – finden Sie Wörterbücher zum Deutschen mit unterschiedlichen inhaltlichen Schwerpunkten (mehr unter Projekt OWID).
Sie können über die allgemeine Stichwortsuche auf alle Wörterbücher dieses Informationssystems zugreifen. Über die „Suche in"-Liste unter dem Suchfenster können Sie festlegen, ob alle in OWID enthaltenen Wörterbücher in die Suche einbezogen werden sollen oder nicht. Darüber hinaus können Sie über „Optionen" auswählen, ob Sie die Suche über die gesamte Stichwortliste von OWID laufen lassen wollen (diese besteht aus etwa 300.000 Stichwörtern, zu denen es in der Mehrzahl Informationen zur Rechtschreibung, in Kürze mit automatisch gewonnenen Belegen gibt), oder ob Sie die Suche auf lexikografisch ausgearbeitete Artikel mit tiefergehenden Informationen aus den einzelnen Wörterbüchern einschränken wollen (mehr Informationen zur Suche).
Wenn Sie sich für eines der Wörterbücher in OWID speziell interessieren, klicken Sie bitte auf die entsprechende Schaltfläche.
elexiko
Neologismenwörterbuch
Feste Wortverbindungen
Diskurswörterbuch 1945-55
© Institut für Deutsche Sprache (http://www.ids-mannheim.de)
Startseite OWID | Projekt OWID

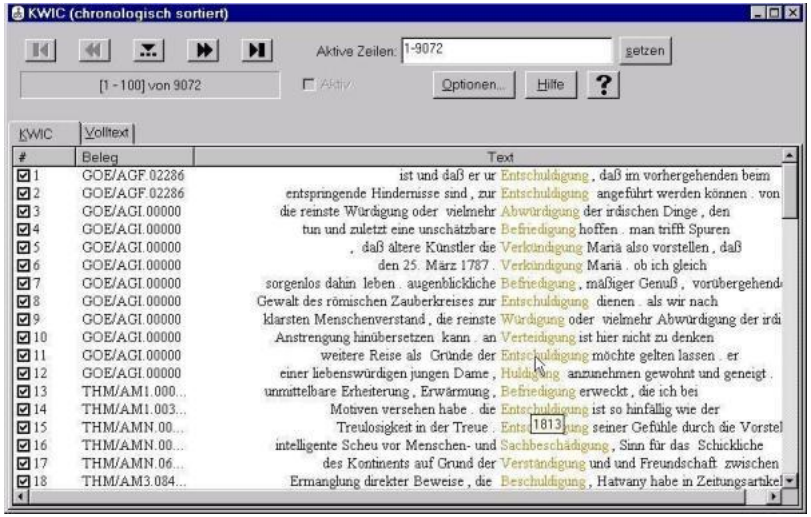
Deutsches Referenzkorpus (DeReKo)

- weltweit größte Sammlung deutschsprachiger Korpora (ca. 5 Milliarden Wörter, umfangreiche Annotationen)
- Texte aus Gegenwart und jüngerer Vergangenheit als Dokumentation des tatsächlichen Gebrauchs der deutschen Sprache in der geschriebenen Modalität
- empirische Basis für nationale und internationale Germanisten, sowie für interdisziplinäre Forschungen
- überwiegend urheberrechtlich geschütztes Material, daher:
- kein Zugriff auf vollständige Korpustexte, sondern nur auf begrenzte Kontexte

COSMAS II


Ergebnisse von 'derartig*' in Korpus: mk - Mannheimer Korpora 1+2
Ansicht Andocken Extras Hilfe
Treffer Texte Dokument
1 1 MK1/LBC Heinrich Böll: Ansichten eines Clowns, [Roman], 1963
8 1 MK1/LFH Max Frisch: Homo faber, [Roman], 1957
1 1 MK1/LGB Günter Grass: Die Blechtrommel, [Roman], 1962
1 1 MK1/LSO Erwin Strittmatter: Ole Bienkopp, [Roman], 1963
4 1 MK1/TPM Heinz Pinkwart: Mord ist schlecht für hohen Blutdruck,
1 1 MK1/WBO Peter Bamm: Ex ovo, [Sachbuch], 1956
3 1 MK1/WGW Otto Willi Gail ; W. Petri: Weltraumfahrt, [Sachbuch],
1 1 MK1/WHK Hermann Heimpel: Kapitulation vor der Geschichte? [Sac
2 1 MK1/WJA Karl Jaspers: Die Atombombe und die Zukuft des Menschei
20 1 MK1/WPE Rudolf Pörtner: Die Erben Roms, [Sachbuch], 1964
24 1 MK1/WUB Karl Ullrich: Wehr Dich, Bürger! [Sachbuch], 1960
8 3 MK1/ZBW Bild der Wissenschaft, [Fachzschr.], Jan. - März 1967
1 1 MK1/ZB4 Bildzeitung, [Tagesztg.], April 1967 (ausgewählte Artil
2 2 MK1/ZB6 Bildzeitung, [Tagesztg.], Juni 1967 (ausgewählte Artik
21 17 MK1/ZFA Frankfurter Allgemeine, [Tagesztg.], Dez. 1965 - Febr.
9 1 MK1/ZSG Studium generale, [Fachzschr.], Dez. 1966, (ausgewählt
5 2 MK1/ZUR Urania, [Fachzschr.], 1966-1967, (ausgewählte Artikel)
3 3 MK1/ZWE Die Welt, [Tagesztg.], Jan. - Febr. 1966, (ausgewählte
3 1 MK2/HAB Anweisungen, (1961, 1966, 1972)
1 1 MK2/PRO Informationsmaterial, (1960-1974)
1 1 MK2/TRI Trivialliteratur, (1960-1974)
1 1 MK2/WF1 Forschungsberichte des Instituts für Deutsche Sprache,
9 7 MK2/ZTG Zeitungen und Zeitschriften, (1960, 1969, 1973)
130 51 23 Dokumente


Textsuche in Korpus: hist - historische Korpora
Korpus Palette Optionen Fenster Hilfe
$ SUCHWORT
$ X NICHT Y
$ LEMMA
$ X ABSTAND Y
Bsp. Mauerspecht
STR( )
'beispiellos*'
beispiellos*
Neu
Suchen
beispiellos*


KWIC (chronologisch sortiert)
Aktive Zeilen: 1-9072
setzen
[1 - 100] von 9072
Optionen...
Hilfe
KWIC
Volltext
Beleg Text
1 GOE/AGF.02286 ist und daß er ur Entschuldigung , daß im vorhergehenden beim
2 GOE/AGF.02286 entspringende Hindernisse sind , zur Entschuldigung angeführt werden können . von
3 GOE/AGI.00000 die reinste Würdigung oder vielmehr Abwürdigung der irdischen Dinge , den
4 GOE/AGI.00000 tun und zuletzt eine unschätzbare Befriedigung hoffen . man trifft Spuren
5 GOE/AGI.00000 , daß ältere Künstler die Verkündigung Maria also vorstellen , daß
6 GOE/AGI.00000 den 25. März 1787 . Verkündigung Maria . ob ich gleich
7 GOE/AGI.00000 sorgenlos dahin leben . augenblickliche Befriedigung , mäßiger Genuß , vorübergehend
8 GOE/AGI.00000 Gewalt des römischen Zauberkreises zur Entschuldigung dienen . als wir nach
9 GOE/AGI.00000 klarsten Menschenverstand , die reinste Würdigung oder vielmehr Abwürdigung der irdi
10 GOE/AGI.00000 Anstrengung hinübersetzen kann . an Verteidigung ist hier nicht zu denken
11 GOE/AGI.00000 weitere Reise als Gründe der Entschuldigung möchte gelten lassen . er
12 GOE/AGI.00000 einer liebenswürdigen jungen Dame , Huldigung anzunehmen gewohnt und geneigt .
13 THM/AM1.000... unmittelbare Erheiterung , Erwärmung , Befriedigung erweckt , die ich bei
14 THM/AM1.003... Motiven versehen habe . die Entschuldigung ist so hinfällig wie der
15 THM/AMN.00... Treulosigkeit in der Treue . Entsc 1813 ung seiner Gefühle durch die Vorstel
16 THM/AMN.00... intelligente Scheu vor Menschen- und Sachbeschädigung , Sinn für das Schickliche
17 THM/AMN.06... des Kontinents auf Grund der Verständigung und und Freundschaft zwischen
18 THM/AM3.084... Ermanglung direkter Beweise , die Beschuldigung , Hatvany habe in Zeitungsartikel

Linguistische Primärdaten zur gesprochenen Sprache

- Archiv für Gesprochenes Deutsch
  - 44 Korpora
  - 6.000 h Audio-Aufzeichnungen
  - 6.000.000 Token transkribiert
- Datenbank für Gesprochenes Deutsch
  - aufbereitete Daten
  - web-basierter Zugang
  - Relaunch 2012 als DGD 2.0
  - Nationales Referenzkorpus

Archiv für Gesprochenes Deutsch (AGD) / Datenbank für Gesprochenes Deutsch (DGD)

- Sammlung von Korpora des gesprochenen Deutsch
- Audioaufnahmen, Videoaufnahmen, Transkriptionen
- Webschnittstelle (DGD)

DGD

DATENBANK FÜR GESPROCHENES DEUTSCH

Herzlich willkommen in der Datenbank für Gesprochenes Deutsch (DGD 2.0 - extern)

Schnelleinstieg

Verschaffen Sie sich über den Menüpunkt "**Über die DGD > Bestand**" einen **Überblick** über die verfügbaren Daten.
Detaillierte **Beschreibungen** der Korpora (Korpusmetadaten) finden Sie unter "**Korpora > Korpusbeschreibungen**".

Zum **Browsen** der Ereignis- oder Sprecherdokumentationen sowie von Transkripten und Zusatzmaterialien bzw. zum **Anhören** von Audioaufnahmen eines ausgewählten Korpus wählen Sie die entsprechenden Unterpunkte im Menü "**Korpora**".

Zum **Durchsuchen** von Ereignis- oder Sprecherdokumentationen sowie von Transkripten wählen Sie die entsprechenden Unterpunkte im Menü "**Recherche > Volltext**".

Einen direkten Zugang zu den **FOLK-Daten** erhalten Sie zum Beispiel, indem Sie die zugehörigen Transkripte über "**Korpora > Transkripte**" browsen oder über "**Recherche > Volltext**" durchsuchen.

Forschungsdatenmanagement

1. Eingabe (Ingest Model)
2. Ausgabe und Zugriff (Access Model)
3. Technische Basis (Repositorium)
4. Nachhaltiger Betrieb (Maintenance Model)

Ingest Model (Input)

- Datenaufbereitung
- Definition von Standards
- lebendes Archiv vs. Langzeitarchiv

Access Model (Output)

- Identifizierbarkeit
- Referenzierbarkeit
- Zugriffsberechtigungen

Repositorium (Technik)

- Phys. Speicherstrategien
- Implementierung Input & Output
  - Ingest Model
  - Access Model
- Langzeitarchivierung

Maintenance Model

- technische Gewährleistung der Nachhaltigkeit
- strukturelle Gewährleistung der Nachhaltigkeit
- Implementierung von Weiterentwicklungen
- Etablierung von Workflows bei Nutzern

Arbeitsabläufe beim Umgang mit Primärdaten

Bedarf:

- Suche in den vorhandenen Daten
- Referenzierung von gefundenen Daten
- Definition von eigenen Zusammenstellungen
- Zugriff auf eigene Definitionen
- Entdeckung von unbekannten Daten

Lösungen

- Nutzung oder Entwicklung von Standards
- Definition von Workflows

Standardisierungsbeispiel 1: Metadaten

Bedarf:

- Anforderungen wie für Primärdaten
- umfassender bzw. universeller Katalog (nicht möglich)
- dynamische Metadaten-Schemata

Lösungen

- Component MetaData Infrastructure (CMDI)
  - komponentenbasierter Aufbau mit Registrierung
  - konform zu ISOcat (ISO 12620, Data Category Registry)
- fester, allgemeingültiger „Kern-Katalog“
- flexible, bedarfsspezifische Ergänzungen
- Harvesting: OAI-PMH-Schnittstelle

Standardisierungsbeispiel 2: Identifizierung von Ressourcen

- Stabile Referenzen durch Persistent Identifier
  - Dienst zum Registrieren und Auflösen von PIDs
  - Verwendung von Handles
  - unabhängig vom Freiheitsgrad der Verfügbarkeit einer Ressource
- Dies erlaubt u.a. Querverweise zwischen Primärdaten und Publikationen

Technisches Bewusstsein

- Softwarekonzept als tragende Säule
- Nachhaltigkeit durch modulare Architekturen
- Kontinuierliche Analyse und Hinterfragung der bestehenden Softwarelösungen

Vernetzung

- Neben der Mitwirkung in der WGL-AG zu Forschungsdaten ist das IDS vielfältig vernetzt:
  - Fachspezifische Vernetzung
  - Fächerübergreifende Vernetzung
- Netzwerke:
  - Verbundprojekte
  - Offizielle Gremien
  - Interessenvereinigungen

Projekte und Mitgliedschaften

FI1

Aufbau eines Zentrums „Digitale Forschungsressourcen für die germanistische Sprachwissenschaft"
2009-2010 – gefördert vom BMBF

FI2

Forschungsinfrastrukturen in wissenschaftlichen Einrichtungen: Implementierung eines Prototyps am Institut für Deutsche Sprache
2011-2013 – gefördert vom BMBF

DSPIN

Deutsche Sprachressourcen-Infrastruktur D-SPIN
2008-2011 – gefördert vom BMBF

CL/D

Common Language Resources and Technology Infrastructure CLARIN in Deutschland
2008-2011 – gefördert vom BMBF

CL/EU

Common Language Resources and Technology Infrastructure CLARIN in Europe
2011-2014 – gefördert vom BMBF

TG

TextGrid – Vernetzte Forschungsumgebung in den eHumanities
2009-2012 – gefördert vom BMBF

WG

WissGrid – Grid für die Wissenschaft
2009-2012 – gefördert vom BMBF

FP1

Zentrum für germanistische Forschungsprimärdaten
2011-2012 – gefördert von der DFG

VG

Verwertung Geist – Analyse und Konzepterstellung zur systemat. Verwertung geistes- und sozialwissensch. Forschungsergebnisse
2011-2013 – gefördert vom BMBF

DIN

Deutsches Institut für Normung, DIN
seit 2004 – ohne Förderung

ISO

Internationale Organisation für Normung, ISO
seit 2004 – ohne Förderung

TEI

Text Encoding Initiative, TEI
seit 2011 – ohne Förderung

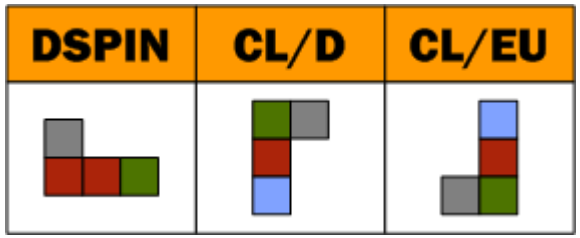
nestor

Network of Expertise in long-term Storage and availability of digital Resources in Germany, nestor
seit 2010 – ohne Förderung

FI2

Forschungsinfrastrukturen in wissenschaftlichen Einrichtungen: Implementierung eines Prototyps am IDS

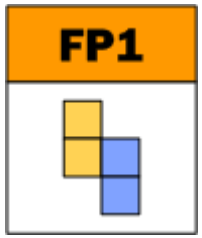
- Prototypische Einrichtung einer Organisationseinheit „Forschungsinfrastrukturen“ am IDS
- Entwicklung einer Art infrastrukturellen Werkzeugkastens
- Verwendung der Arbeitsergebnisse als mögliche Blaupause für andere Forschungsinstitute, die institutionelle Lösungen für den Umgang mit Forschungsdaten suchen



Deutsche Sprachressourcen-Infrastruktur; Common Language Resources and Technology Infrastructure in Deutschland / Europe

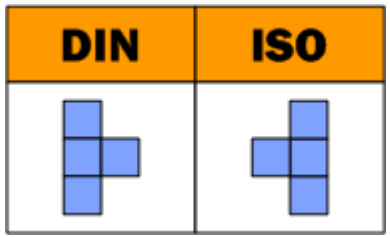
- Paneuropäische Initiative zum koordinierten Umgang mit Sprachressourcen und –technologien (ERIC)
- Organisation, Ausbau, Vernetzung und Vereinheitlichung existierender nationaler Infrastrukturen, EU-geförderter Projekte und ihrer Ressourcen
- Aufbau von dedizierten Zentren für eine funktionsfähige Infrastruktur
- Ressourcenanbieter-Föderation innerhalb des DFN
- Darstellung der rechtlichen Rahmenbedingungen
- Bereitstellung von Ressourcen, Daten und Werkzeugen über State-of-the-Art-Registraturen und Web-Services

CLARIN-D



Zentrum für germanistische Forschungsprimärdaten

- Etablierung einer einheitlichen Infrastruktur für die nachhaltige Bereitstellung von Forschungsprimärdaten aus der germanistischen Linguistik
- Entwicklung datentechnischer, dokumentatorischer und rechtlicher Standards sowie Best-Practice-Richtlinien für die nachhaltige Archivierung und Verfügbarmachung von sprach-wissenschaftlichen Primärdaten
- Perspektive: Öffnung des Zentrums für germanistische Forschungsdaten fremder Institutionen



Deutsches Institut für Normung, DIN; Internationale Organisation für Normung, ISO

- nationale und internationale Vereinigung von Normungsorganisationen
- Mitarbeit im ISO-Komitee ISO/TC37/SC4 (Terminology and Other Language and Content Resources: Language Resources Management) und in der nationalen Spiegelgruppe in DIN

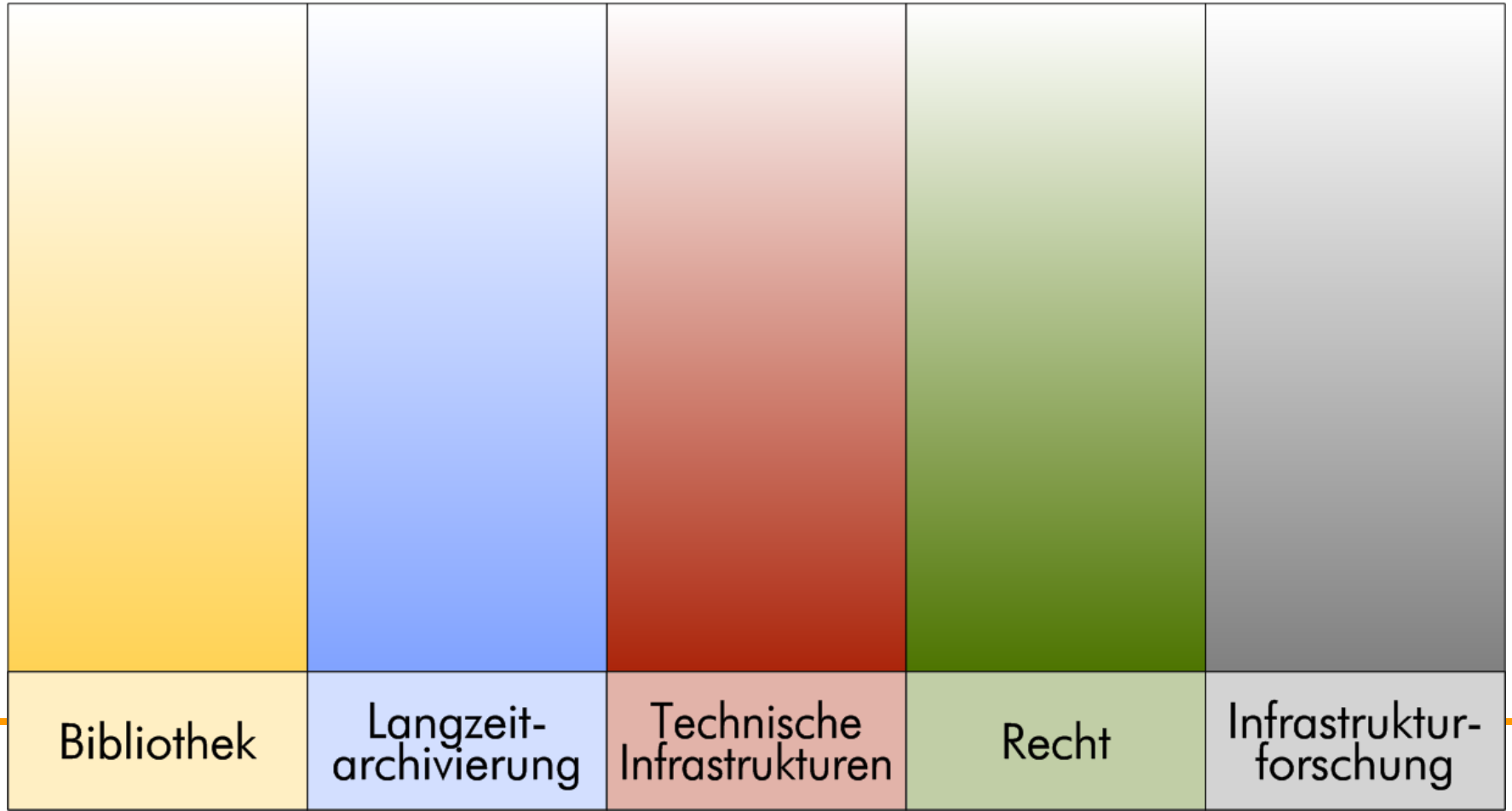
ISO DIN

Network of Expertise in long-term Storage and availability of digital Resources in Germany, nestor

- Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen in Deutschland
- IDS ist Partner seit 2010

nestor

Programmbereich Forschungsinfrastrukturen



Forschungsdaten in den Geisteswissenschaften – die germanistische Sprachwissenschaft

Andreas Witt

Institut für Deutsche Sprache, Mannheim

Workshop Forschungsdaten
WGL Geschäftsstelle Berlin
2012-05-10