

SICHTWEISE VON CLARIN

ZUKÜNFTIGE AUFGABEN,
KOMPONENTEN,
HERAUSFORDERUNGEN
THORSTEN TRIPPEL, UNIVERSITÄT TÜBINGEN

- Zweites European Research Infrastructure Consortium (ERIC)
- Mitglieder:
 - Bulgarien
 - Dänemark
 - Deutschland
 - Estland
 - Griechenland
 - Litauen
 - Niederlande
 - Niederländische Sprachunion (Belgien, Niederlande)
 - Polen
 - Österreich
 - Schweden
 - Tschechische Republik
- Beobachter:
 - Norwegen
- Kandidaten:
 - Finnland
 - Großbritannien

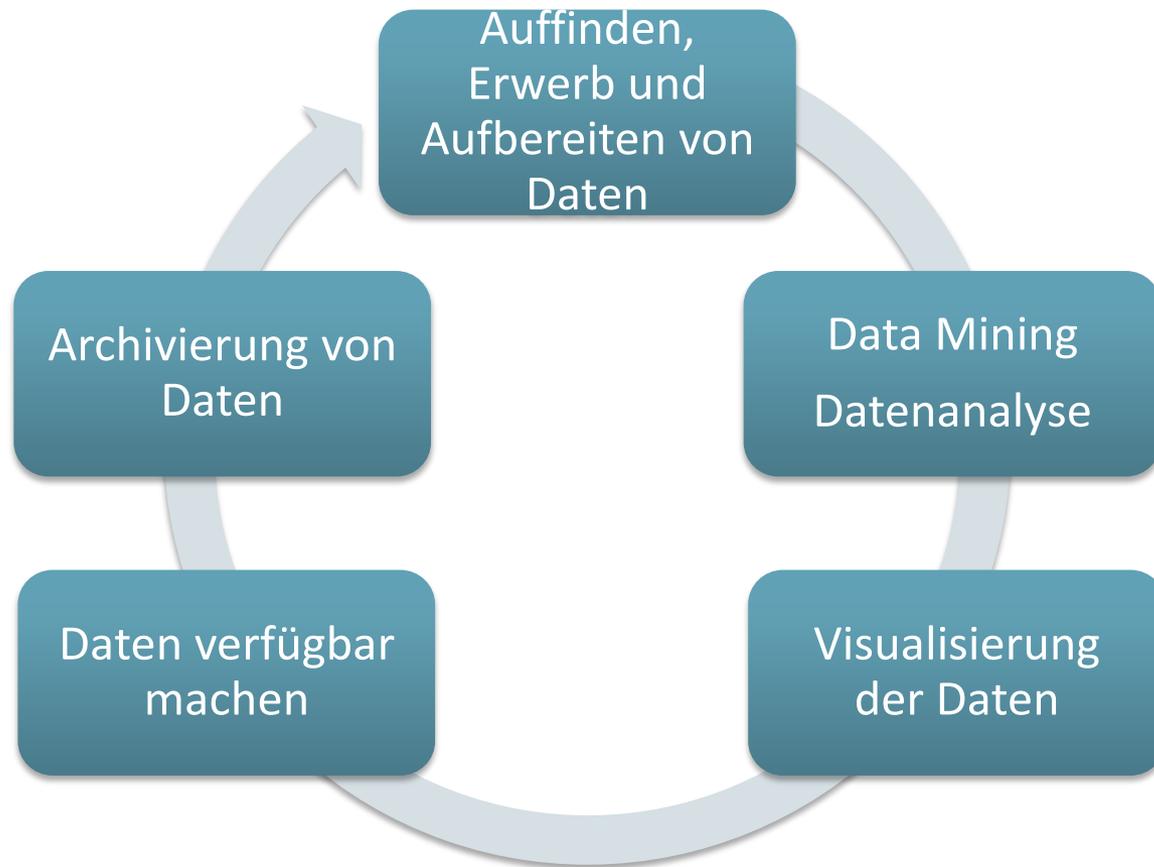


CLARIN in Deutschland

- Zentren (Repositorien und Dienste für die Community):
 - Berlin: BBAW
 - Nijmegen: MPI
 - Mannheim: IDS
 - Universität Hamburg
 - Universität Leipzig
 - Universität München (LMU)
 - Universität des Saarlands
 - Universität Stuttgart
 - Universität Tübingen
- Rechenzentren(E):
 - Garching: RZG
 - Göttingen: GWDG
 - Jülich: RZJ
- Facharbeitsgruppen(F-AGs)
 - 10 F-AGs mit jeweils 10-20 Mitgliedern (nicht dargestellt)
 - Geistes- und Sozialwissenschaften



- Einfacher Zugang zu Sprachdaten mit einheitlichem Login (Single Sign-On)
- Verteilte Dienste für die Forschung
- Archivierung und nachhaltiger Zugang zu Sprachdaten





Konference KL2014

Krátké zprávy

Co je korpus?
Kontakty
Dostupné korpusy
Projekt InterCorp
Naše publikace
Dohody a registrace

Hledat v ČNK

Pracovní kolektiv
Poskytovatelé textů
Granty, sponzoři
Jak citovat korpus
Kritikova
Výuka
Spolupráce
Jiné korpusy
Ke stažení
Manuál a instalace



Ústav Českého národního korpusu

Filozofická fakulta Univerzity Karlovy v Praze

Hledat na stránkách

Portál ČNK

Od začátku roku 2014 funguje nový portál pro práci s korpusy na adrese www.korpus.cz. V rámci této stránky najdete nejen všechny naše nástroje pro práci s korpusovými daty, zejména pak korpusový vyhledávač **KonText**, nástroj pro vyhodnocování variant **SyD** a **dášíř**, ale také manuál ve formě **Wiki** a **poradno** pro uživatele.

Co je korpus?

Korpus je soubor počítačově uložených textů (v případě mluveného jazyka - přepisů záznamu mluvy), který primárně slouží k jazykovému výzkumu. K práci s korpusy slouží speciální vyhledávací program. S jeho pomocí je možné vyhledávat slova a slovní spojení v kontextu a zjistit jejich frekvenci v korpusu i původní textový zdroj. Umožňuje i další zpracování nalezeného (např. abecední třídění apod.). U některých korpusů lze vyhledávat i podle slovních druhů.

Český národní korpus (ČNK) je akademický projekt zaměřený na budování rozsáhlého počítačového korpusu především psané češtiny. Pracuje na něm **Ústav Českého národního korpusu** na Filozofické fakultě Univerzity Karlovy v Praze (ÚČNK). Od svého založení roku 1994 má ÚČNK na starosti budování

Aktuality

- Korpusy Aranea**
V květnu 2014 byla v ČNK zveřejněna rodina nereferenčních srovnávacích webových korpusů Aranea sestavených Vladimírem Benkem. V současné době pokrývá celkem 7 jazyků (de, en, fr, nl, ru, sk), další informace [zde](#).
- Proverbs: Their Lexical and Semantic Features**
Nedávno vyšla v USA anglicky psaná Františka Čermáka o paremiologii (přísluvnicví) pojednávající o různých, hlavně však lexikálních aspektech více jazyků včetně jejich pragmatiky, paremiologických míním (češtiny, angličtiny a němčiny). Je založená na rozsáhlé analýze dat, většinou korpusových. K nám nedostupnou, lze v omezeném rozprohově zájemce získat přímo od autora. Více [zde](#).

AUSTRIAN ACADEMY CORPUS



Institution

Text-Technology Applications Lab

Austrian Academy Corpus

Das **AAC-Austrian Academy Corpus** ist eine umfangreiche und komplex strukturierte Sammlung von digitalen Volltexten zur deutschen Sprache und Literatur im Untersuchungszeitraum 1848 bis 1989. Das AAC ist sowohl Basis für die Entwicklung innovativer texttechnologischer Anwendungen als auch für die Verfolgung philologischer und kulturwissenschaftlicher Forschungsinteressen. Das Programm erhebt den Anspruch, Schnittstellen zwischen kulturwissenschaftlichen Methoden und Prinzipien (u.a. philologische Genauigkeit, Editionstechniken, Textstrukturierung, Annotierung) und IKT-Anwendungen zu schaffen sowie Schnittstellen zwischen Grundlagenforschung und deren Umsetzungsmöglichkeiten.

Im AAC werden Texte digitalisiert, um sie philologisch zu bearbeiten, auszuwerten und zu interpretieren. Das Erkenntnisinteresse des AAC gilt den immer neu zu stellenden Fragen der Philologie: Wie funktionieren Wörter und Texte, woher kommen sie, wie sind sie gebaut, welche Inhalte vermitteln sie, warum werden sie geschrieben, warum werden sie gelesen, welchen Einfluss haben sie auf den Sprachgebrauch? Die EDV und texttechnologische Methoden machen es heute möglich, diese Fragen systematisch an äußerst große Mengen von Texten heranzutragen und sie in einem Team beschleunigt abzuarbeiten; durch diese geänderte Methodik in den Textwissenschaften gibt es neue Hypothesen und neue Antworten und das wiederum bringt Spannung, Kontroverse und neue Erkenntnisse in die Philologie.





INSTITUT FÜR DEUTSCHE SPRACHE

Mitglied der Leibniz-Gemeinschaft

Aktuelles | **Organisationsstruktur** | **Forschung** | **Bibliothek** | **Service** | **Publikationen** | **Über uns**

Direktion und zentrale Forschung | **Grammatik** | **Lexik** | **Pragmatik** | **Öffentlichkeitsarbeit** | **Zentrale DV-Dienste** | **Verwaltung**

Startseite : : Organisationsstruktur : : **Direktion** : : Korpuslinguistik | Korpuslinguistik : : Projekte : : **Korpusausbau**

Direktion und zentrale Forschung

- : : Forschungsinfrastrukturen
- : : Korpuslinguistik
- : : Projekte
- : : **Korpusausbau**
 - : : Einsatzgebiete
 - : : Akquisition
 - : : IDS-Textmodell
 - : : Archiv
 - : : Ausgaben
 - : : Annotationen
 - : : Verfügbarkeit
 - : : Textgeber
 - : : Ehemalige Mitarbeiter
 - : : Kooperationen
 - : : Bibliographie (IDS)
 - : : Bibliographie
 - : : Analysemethoden
 - : : Kredite

Ausbau und Pflege der Korpora geschriebener Gegenwartssprache

Kohlhammer

DAS DEUTSCHE REFERENZKORPUS – DEREKO

die weltweit größte Sammlung deutschsprachiger Korpora als empirische Basis für die linguistische Forschung

Die Korpora geschriebener Gegenwartssprache des IDS

- bilden mit 6,1 Milliarden Wörtern (Stand 19.03.2013) die weltweit größte linguistisch motivierte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit.
- enthalten belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Zahl von Zeitungstexten sowie eine breite Palette weiterer Textarten und werden kontinuierlich weiterentwickelt.
- werden im Hinblick auf Umfang, Variabilität, Qualität und Aktualität akquiriert und erlauben in der Nutzungsphase über COSMAS II die Komposition virtueller Korpora, die repräsentativ oder auf spezielle Aufgabenstellungen zugeschnitten sind.
- enthalten ausschließlich urheberrechtlich abgesichertes Material.

Kontakt:
<korpuslinguistik@ids->

Leitung:
Dr. Marc Kupietz
<kupietz@ids->

Wissenschaftliche Mitarbeiter:
Cyril Belica <belica@ids->
Dr. Harald Lungen
<lungen@ids->
Rainer Perkuhn
<perkuhn@ids->

Kooperationen:
siehe hier

Ehemalige am Korpusaufbau beteiligte Mitarbeiter des IDS:
siehe hier

KEELEVEEB

Otsi sõna:

Sõnastikud

1. Klõpsa alltoodud loendis sõnastikel, kust soovid otsida.
2. Kirjuta lehekülje päises olevasse lahtrisse otsisõna ja vajuta Enter-klahvi. Keeleveeb teeb päringu kõikidesse valitud sõnastikesse korraga.
3. Aktiivse sõnastiku vahetamiseks klõpsa hallidel vahekaartidel.
4. Sõnastiku pealehe vaatamiseks tee päring tühja otsisõnaga.

Näita sõnastikku kohe

Eesti keele sõnastikud

- Õigekeelsussõnaraamat 2013 (EKI)
- Eesti Entsüklopeedia (MTÜ Entsüklopeedia)
- Tesaaurus (Filosoft)
- Sünonüümisõnastik (IBS)
- Eesti keele seletav sõnaraamat (EKI)
- Eesti Wordnet ehk TEKsaurus (TÜ)
- Kõnekäändude ja fraseoloogismide andmebaas (KM)
- Eesti etümoloogiasõnaraamat (EKI)
- Kõnekäändud ja fraseoloogismid (KM)
- Eesti vanasõnad (KM)
- Esimene Eesti Slängi Sõnaraamat (EKI)

Tõlkesõnastikud

- Inglise<->eesti (Na)
- Inglise<->eesti (Aa)
- Inglise<->eesti (Fild)
- Inglise<->eesti (IBS)
- Vene-eesti (EKI)
- Eesti-vene (EKI)
- Soome<->eesti (IS)
- Rootsi<->eesti (IS)
- Norra<->eesti (EKI)
- Eesti<->läti (Tilde)
- Võru<->eesti (VI)

D W D S

Ressourcen | Erschließung | Projekt | Aktuelles

Suche im DWDS

Beispiele: Balkon, Experte, Verlierer

Das Digitale Wörterbuch der deutschen Sprache

Ein Wortsuchsystem zur deutschen Sprache in Geschichte und Gegenwart: Hintergrund

- über 410.000 lexikografisch geprüfte Einträge aus fünf Wörterbuchquellen: Wörterbücher
- 1,8 Milliarden Korpusbelege aus 15 Korpora: Korpora
- statistische Wortprofile und Wortverlaufskurven: Statistiken

Neuigkeiten

DWDS und DTA mit einem Beitrag zu historischen Zeitungen heute auf der IREC 2014 in Reykjavik: [bit.ly/1uKwU](#) #irec #zotling 26. Mai

RT @usbotschaft: Junge Amerikaner wählen ihr deutsches #Lieblingswort: [ow.ly/x395d](#) 21. Mai

[Alle Neuigkeiten auf Twitter](#)

DWDS - ein Projekt der BBAW

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

De Geïntegreerde Taal-Bank

Basiszoeken | Uitgebreid zoeken | Zoeken in bronnenlijst | Resultaat | Help & Info

Zoek in de woordenboeken:
 ONW VMNW MNW WNT WFT

Mod. Ned. trefwoord

Origineel trefwoord

Woordsoort

Woord in betekenis

Woord in citaat

Woord in artikel

Start zoeken | Wis velden | Geschiedenis | Symbool | H&A | ?

Versie 1.3, 2 juli 2010

Nieuw in deze versie

- Suche nach Ressourcen
- Verwendung der Metadatenätze
- Automatische Verarbeitung von Daten aus zugänglichen Quellen (OAI-PMH Provider für CMDI)



The screenshot shows the VLO search results page. At the top, there is a header with the VLO logo and the tagline "Explore the world of language resources and technology from different perspectives". Below the header, there is a search bar and a "Search" button. The main content area displays a list of search results, each with a title, a brief description, and an "Expand" button. The results include:

- !De kal Lama** (http://www.sil.org/resources/archives/35552)
- !Kal n'n su len** (2nd ed. (1992), 18 pp.; http://www.sil.org/resources/archives/35390)
- !Kal n'n wonke** (repub. (1992), 18 pp.; http://www.sil.org/resources/archives/35345)
- !Komokenzj eme phé** (Adapted from Æsop's fables; http://www.sil.org/resources/archives/36180)
- !O!ung Swadesh List** (Resources: | 1 other |)
- !O!ung: a language of Angola** (A page from the Web edition of Ethnologue: Languages of the World (15th edition) giving basic facts about the language)

On the right side of the page, there is a "NARROW DOWN" section with a search input field and a list of filters:

- LANGUAGE
- COLLECTION
- RESOURCE TYPE
- CONTINENT
- COUNTRY
- MODALITY
- GENRE



Zadejte slovo nebo slovní spojení

nerozlišovat velikost písmen vyhledávat slovní tvary zadaného **lemmatu**

Korpusový vyhledávač **konText**

Aplikace **konText** představuje základní **vyhledávací program pro práci s korpusy** projektu ČNK (nahrazuje tak původní **NOŠKE**). Umožňuje tvorbu a vyhodnocení jednoduchých i komplexních dotazů, zobrazení formou konkordančních řádků, vypočtení frekvenční distribuce, asocičních měr kolokaci a další práci s jazykovými daty. Přehledné jsou všechny funkce konTextu popsány v [manuálu](#).

Jak to vlastně je?

Jak mluvíme? Vlastně jako ani moc nevíme? Podívejme se na časté výpítkové slovíčko vlastně do korpusu mluveného jazyka a srovnáme, jak se například liší mluva mladších a starších. Více se dozvíte v [manuálu](#) práce s korpusem.

V korpusu ORAL2008 vyberte typ dotazu **slovní tvar**, zadejte:

Ústav Českého národního korpusu

Český národní korpus je akademický projekt založený v roce 1993 Ústavem Českého národního korpusu. Jeho cílem je mapovat vývoj češtiny (psané i mluvené) ve všech jejích variantách. Korpusy ČNK jsou po **bezplatné registraci** otevřeny všem uživatelům. Pokud touží vědět, jak se čeština používá - [více informací](#).

Jazykový korpus je rozsáhlý soubor autentických textů, které jsou pomocí speciálních nástrojů vyhledávány jazykové jevy (např. slova a fráze) v reálných datech podloženy jazykový výzkum v rozsahu, který je pro výzkum potřebný.

Dostupné korpusy

Projekt Český národní korpus nabízí uživatelům mnoho jazykových typů výzkumů. K dispozici jsou korpusy **psaného** jazyka, **psaného** jazyka, **paralelní** korpus **InterCorp** s více než **30 jazyky** a **specializovaných**. Jejich přehled s podrobnějšími informacemi najdete [zde](#).

DTA

In den Titelziten im Korpus in der Dokumentation

Ranke, Leopold von: Die römischen Päpste. Bd. 2. Berlin, 1836.

Bild: << vorherige Seite

bündnerischen Päffen bis an den Comersee, eingenommen: 35000 Mann stark flog alsdann dieses Heer längs der Adda und dem Oglio hinab. Noch einmal ward der Herzog von Mantua aufgefordert sich zu unterwerfen. Er erklärte, er stehe im Schutze des Königs von Frankreich, mit diesem müsse man unterhandeln. Indem nun die Deutschen sich gegen Mantua, die Spanier sich gegen Montserrat bewegten, erschienen auch die Franzosen zum zweiten Male. Sie machten auch dieß Mal Fortschritte; sie nahmen Saluzzo, Pinerolo: aber in der Hauptsache richteten sie nichts aus; nicht einmal den Herzog von Savoyen vermochten sie auf neue zu ihrem Willen zu nöthigen. Die Spanier begannen Cafale, die Deutschen nach kurzem Stillstand Mantua zu belagern: *) sie hatten bei weitem das Uebergewicht.

Kein Wunder, wenn in dieser Lage der Dinge jetzt in **Wien** selbst Erinnerungen an die alte kaiserliche Hoheit laut wurden.

„Man werde den Italienern zeigen, daß es noch einen Kaiser gebe, man werde Rechnung mit ihnen halten.“

„Man werde den Italienern zeigen, daß es noch einen Kaiser gebe, man werde Rechnung mit ihnen halten.“

Befonders hatte sich Venedig den Haß des Hauses Oestreich zugezogen. Man urtheilte zu **Wien**, daß wenn Mantua einmal gefallen, auch die Terra ferma von Venedig nicht mehr widerstehn könne. In ein paar Monaten müßte man sie haben, dann könne man die kaiserlichen Lehen zurückfordern. Der spanische Gefandte ging noch weiter. Er verglich die spanisch-österreichische Macht mit der römischen.

*) Das alte Buch dell'istoria di Pietro Giov. Copriata erörtert die einzelnen Momente dieser Ereigniß.

Suomi | Ruotsi | Muut kielet | Rinnakaistekst

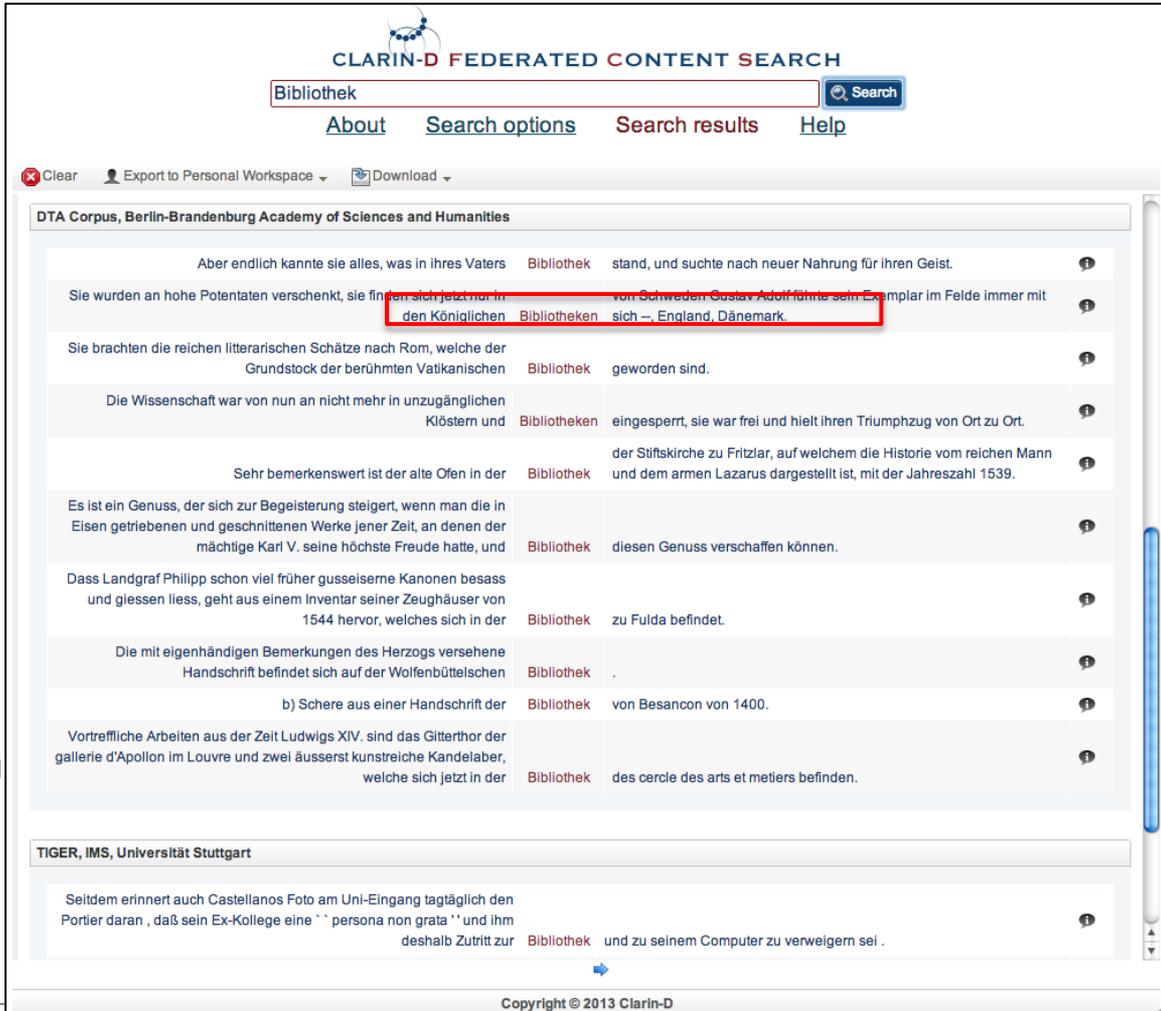
KORP

120 korpusa valittuina — 4 814 427 561 sanetta

Etsi myös alkuosaa loppuosaa ja samaista-pien-ja suuraakoset

Konkordanssi: Tilastoja:

- Verteilte Suche über eine gemeinsame Oberfläche
- Kein zentraler Index
 - Nutzungsrechte
 - Verwendung bestehender Suchdienste
- Webservice basiert
 - Rest-basierte Webservices
 - Kette von Werkzeugen in AAI integriert (kein eigenes Login)



CLARIN-D FEDERATED CONTENT SEARCH

Bibliothek

[About](#) [Search options](#) [Search results](#) [Help](#)

Clear Export to Personal Workspace Download

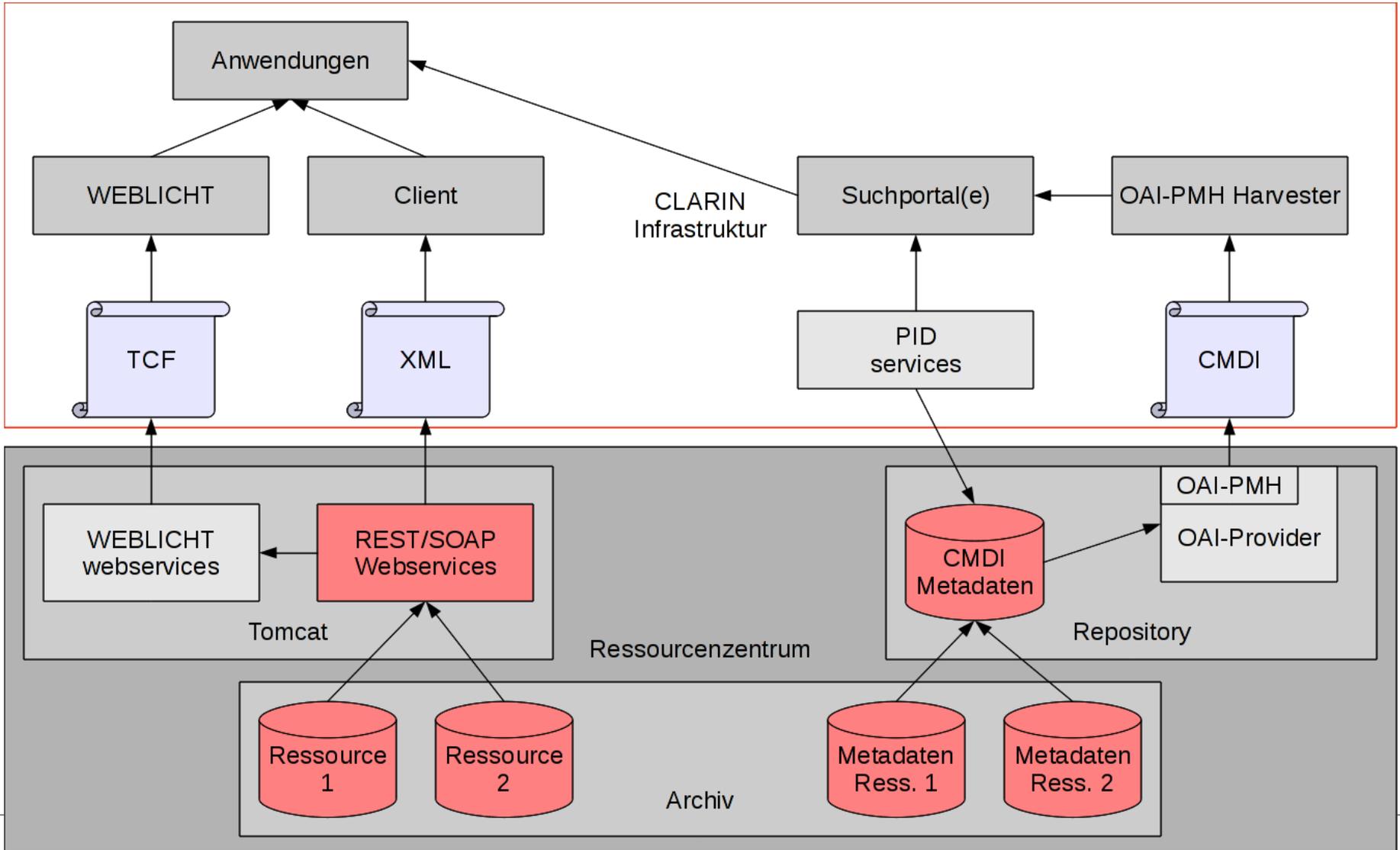
DTA Corpus, Berlin-Brandenburg Academy of Sciences and Humanities

Aber endlich kannte sie alles, was in Ihres Vaters	Bibliothek	stand, und suchte nach neuer Nahrung für ihren Geist.
Sie wurden an hohe Potentaten verschenkt, sie finden sich jetzt nur in den Königlichen Bibliotheken von Schweden, Sussar Adolfführte sein Exemplar im Felde immer mit sich - England, Dänemark.	Bibliothek	
Sie brachten die reichen litterarischen Schätze nach Rom, welche der Grundstock der berühmten Vatikanischen	Bibliothek	geworden sind.
Die Wissenschaft war von nun an nicht mehr in unzugänglichen Klöstern und	Bibliotheken	eingesperrt, sie war frei und hielt ihren Triumphzug von Ort zu Ort.
Sehr bemerkenswert ist der alte Ofen in der	Bibliothek	der Stiftskirche zu Fritzlär, auf welchem die Historie vom reichen Mann und dem armen Lazarus dargestellt ist, mit der Jahreszahl 1539.
Es ist ein Genuss, der sich zur Begeisterung steigert, wenn man die in Eisen getriebenen und geschnittenen Werke jener Zeit, an denen der mächtige Karl V. seine höchste Freude hatte, und	Bibliothek	diesen Genuss verschaffen können.
Dass Landgraf Philipp schon viel früher gusseiserne Kanonen besass und glessen liess, geht aus einem Inventar seiner Zeughäuser von 1544 hervor, welches sich in der	Bibliothek	zu Fulda befindet.
Die mit eigenhändigen Bemerkungen des Herzogs versehene Handschrift befindet sich auf der Wolfenbüttelschen	Bibliothek	.
b) Schere aus einer Handschrift der	Bibliothek	von Besancon von 1400.
Vortreffliche Arbeiten aus der Zeit Ludwigs XIV. sind das Gitterthor der gallerie d'Apollon im Louvre und zwei äusserst kunstreiche Kandelaber, welche sich jetzt in der	Bibliothek	des cercle des arts et metiers befinden.

TIGER, IMS, Universität Stuttgart

Seitdem erinnert auch Castellanos Foto am Uni-Eingang tagtäglich den Portier daran, daß sein Ex-Kollege eine `` persona non grata `` und ihm deshalb Zutritt zur Bibliothek und zu seinem Computer zu verweigern sei.

Copyright © 2013 Clarin-D



- Repository für den Bereich der Psycholinguistik, Kognitionspsychologie und Korpuslinguistik
- Gemeinsamkeit
 - Statistische Analysen
 - Granularität
- Paper Package
 - Experimentrohdaten: Serie von Experimenten
 - Umfragen
 - Zeitgesteuerte Experimente
 - Eyemovement
 - Auswertungsskripte in R
 - Publikation zur Experimentserie
 - i.d.R in LaTeX
 - Eingebette R-Analysen mit KNITR/SWEAVE
 - Erlaubt Reproduzierbarkeit der Analyse (nicht der Rohdaten)
 - Metadaten zum Experiment

- Integration der Metadatenerstellung in den Workflow
 - Beim Archivieren Pflicht
 - Informationen verfügbar? Spätere Ergänzung? Unterstützung?
- PID: Zitation im Paper
 - Verknüpfung von Paper Package und Paper
 - Bibliometrische Verfahren
- Zeitpunkt der Veröffentlichung
 - Nach Verlagsveröffentlichung
 - Nach Review
- Verbindung zu Datenmanagementplänen
 - Generatoren für DMP
 - Aufbau eines Workflows bei den Archiven auf Grundlage der DMP

- Metadaten:
 - Datenkategorien zentral definiert zur semantischen Interoperabilität nach ISO 12620
 - Komponentenmodell für Metadaten nach ISO 24622-1
- Bereitstellung der Daten
 - Via OAI-PMH
 - Verwendung in Spezialsuchmaschine: VLO
- PID:
 - Verknüpfung von Publikation und Rohdaten
 - Reproduzierbarkeit von Analysen durch Ablage der Analyseskripte
- Qualitätssicherung:
 - Über GND (Namen + Institutionen)
 - Beschreibungen