Das CLARIN-D Servicezentrum des Zentrum Sprache an der BBAW

Kai Zimmer
BBAW

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Das Repositorium

Das Repositorium des CLARIN-Servicezentrum des Zentrum Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW), http://clarin.bbaw.de/, dient der Langzeitarchivierung von sprachwissenschaftlichen Primärdaten und konzentriert sich dabei vorwiegend auf historische Textkorpora und lexikalische Ressourcen.

Zertifizierungen

- Es wurde im Juni 2013 als „CLARIN Centre B“ zertifiziert. Siehe das Zertifikat: http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-93
- Die Zertifizierung durch das „Data Seal of Approval“ umfasst insbesondere die Aspekte der Organisation, des Workflows, der Qualitätssicherung und Nachhaltigkeit der Daten. Siehe dazu http://www.datasealofapproval.org

Clarin Repository Anforderungen

- **S**ingle **S**ign**O**n per **A**uthentifizierungs- und **A**uthorisierungs **I**nfrastruktur (SSO/AAI)
- **F**öderierte (Volltext-)**S**uche (FCS)
- **K**omponenten **M**eta**d**aten **I**nfrastruktur (CMDI)
- **P**ersistente **Id**entifizierer (PIDs vgl. DOI)
- **O**pen **A**rchives **I**nitiative/ **P**rotocol für **M**etadata **H**arvesting (OAI/PMH)-Schnittstelle zum Datenaustausch
- Zertifizierung durch Data Seal of Approval

Data Seal of Approval Anforderungen

The Guidelines 2014-2015

Guidelines Relating to Data Producers:

1. The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with disciplinary and ethical norms.
2. The data producer provides the data in formats recommended by the data repository.
3. The data producer provides the data together with the metadata requested by the data repository.

Guidelines Related to Repositories:

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.
5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.
6. The data repository applies documented processes and procedures for managing data storage.
7. The data repository has a plan for long-term preservation of its digital assets.
8. Archiving takes place according to explicit work flows across the data life cycle.
9. The data repository assumes responsibility from the data producers for access and availability of the digital objects.
10. The data repository enables the users to discover and use the data and refer to them in a persistent way.
11. The data repository ensures the integrity of the digital objects and the metadata.
12. The data repository ensures the authenticity of the digital objects and the metadata.
13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

Guidelines Related to Data Consumers:

14. The data consumer complies with access regulations set by the data repository.
15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.
16. The data consumer respects the applicable licences of the data repository regarding the use of the data.

Metadaten und PIDs

- Sämtliche Metadaten im Repositorium sind entsprechend dem CLARIN-spezifischen CMDI-Profil kodiert. CMDI: Component Metadata Infrastructure. Siehe dazu http://www.clarin.eu/content/component-metadata.

- Die einzelnen Datensätze werden mit persistenten Identifizierern (Handle-PIDs) versehen, wodurch eine langfristige Referenzierbarkeit ermöglicht wird. Jede neue Version eines Datensatzes wird mit einer eigenen PID versehen, während die jeweils älteren Versionen über ihre ursprüngliche PID verfügbar bleiben.

Software

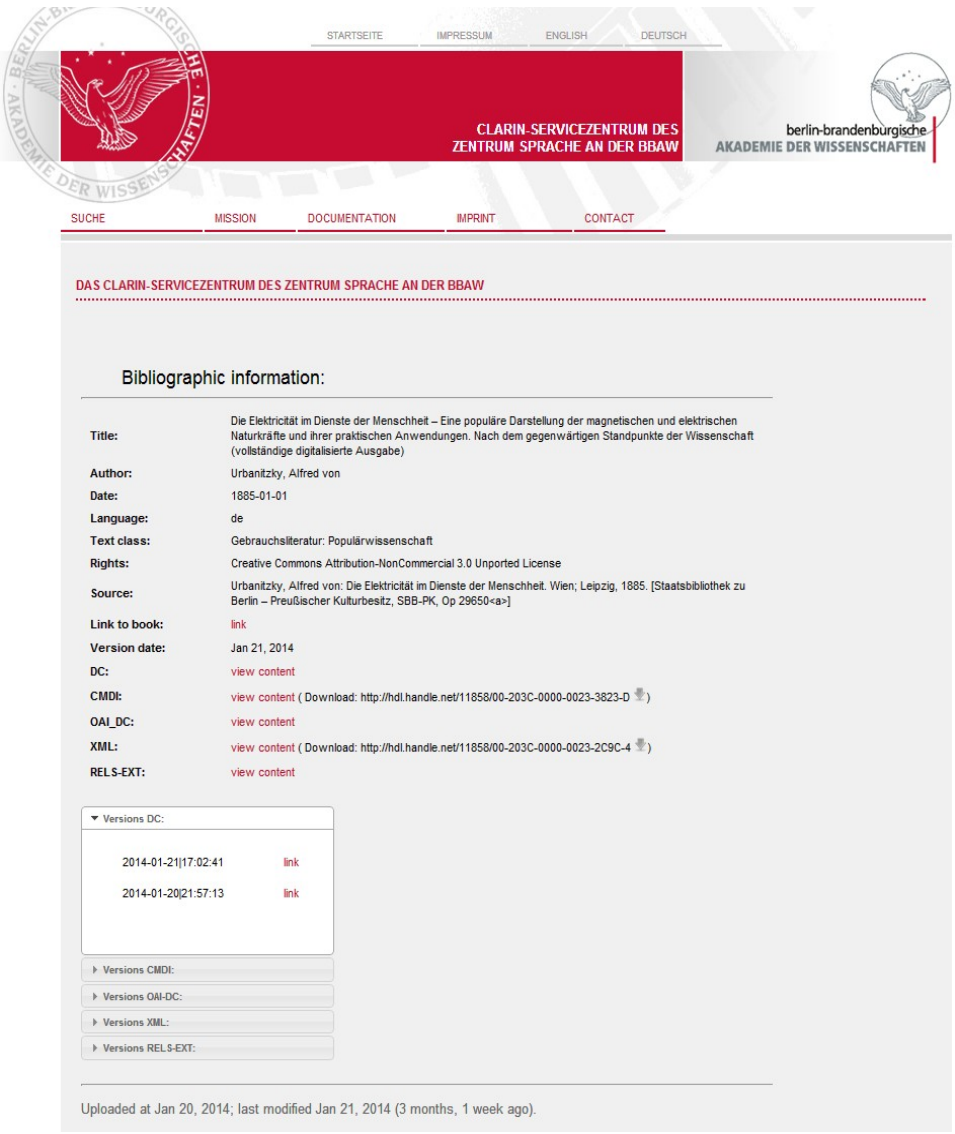
- Das Repositorium basiert, wie die meisten anderen deutschen CLARIN-Repositorien, auf der Software Fedora Commons.
- http://www.fedora-commons.org

- Die Weboberfläche wurde mithilfe des Python-Web-Frameworks Django (http://www.django-de.org/) bzw. der Eulfedora-Komponente (https://github.com/emory-libraries/eulfedora) implementiert, die eine gegenüber der reinen Fedora Commons Software intuitivere Suchmaske und Abfragesyntax ermöglicht.

Suchmaske


STARTSEITE
IMPRESSUM
ENGLISH
DEUTSCH
CLARIN-SERVICEZENTRUM DES ZENTRUM SPRACHE AN DER BBAW
berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN
SUCHE
MISSION
DOCUMENTATION
IMPRINT
CONTACT
DAS CLARIN-SERVICEZENTRUM DES ZENTRUM SPRACHE AN DER BBAW
Author:
Urbanitzky
Title:
Elektricität
Text class:
Bibliographical information:
Date:
between
example: 1600-01-01 or 1800-01-01 - 2013-12-31
Results per page:
10
Results source:
Deutsches Textarchiv
Result fields to display:
Pid
Author
Title
Type
Source
Date
Publisher
Text class
Search
Simple search
1 total objects
Number
Author
Title
Type
Bibliographical information
Date
Publisher
Text class
dta:1216
Urbanitzky, Alfred von
Die Elektricität im Dienste der Menschheit – Eine populäre Darstellung der magnetischen und elektrischen Naturkräfte und ihrer praktischen Anwendungen. Nach dem gegenwärtigen Standpunkte der Wissenschaft (vollständige digitalisierte Ausgabe)
Text
Urbanitzky, Alfred von: Die Elektricität im Dienste der Menschheit. Wien; Leipzig, 1885. [Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, SBB-PK, Op 29650<a>]
1885-01-01
Deutsches Textarchiv
Gebrauchsliteratur: Populärwissenschaft
ZUM SEITENANFANG
CLARIN CENTRE B
hdl enabled
© 2013 BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Metadaten


STARTSEITE
IMPRESSUM
ENGLISH
DEUTSCH
CLARIN-SERVICEZENTRUM DES ZENTRUM SPRACHE AN DER BBAW
berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN
SUCHE
MISSION
DOCUMENTATION
IMPRINT
CONTACT
DAS CLARIN-SERVICEZENTRUM DES ZENTRUM SPRACHE AN DER BBAW
Bibliographic information:
Title: Die Elektricität im Dienste der Menschheit – Eine populäre Darstellung der magnetischen und elektrischen Naturkräfte und ihrer praktischen Anwendungen. Nach dem gegenwärtigen Standpunkte der Wissenschaft (vollständige digitalisierte Ausgabe)
Author: Urbanitzky, Alfred von
Date: 1885-01-01
Language: de
Text class: Gebrauchsliteratur: Populärwissenschaft
Rights: Creative Commons Attribution-NonCommercial 3.0 Unported License
Source: Urbanitzky, Alfred von: Die Elektricität im Dienste der Menschheit. Wien; Leipzig, 1885. [Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, SBB-PK, Op 29650<a>]
Link to book: link
Version date: Jan 21, 2014
DC: view content
CMDI: view content (Download: http://hdl.handle.net/11858/00-203C-0000-0023-3823-D)
OAI_DC: view content
XML: view content (Download: http://hdl.handle.net/11858/00-203C-0000-0023-2C9C-4)
RELS-EXT: view content
Versions DC:
2014-01-21|17:02:41 link
2014-01-20|21:57:13 link
Versions CMDI:
Versions OAI-DC:
Versions XML:
Versions RELS-EXT:
Uploaded at Jan 20, 2014; last modified Jan 21, 2014 (3 months, 1 week ago).

OAI/PMH

Open **A**rchives **I**nitiative **P**rotocol for **M**etadata **H**arvesting (OAI/PMH):
Für den automatisierten Zugriff besitzt das Repositorium eine OAI-PMH-Schnittstelle über die DC- und CMDI-Metadaten ausgeliefert werden.
URL zur Schnittstelle des CLARIN-Repositorys der BBAW:
http://clarin.bbaw.de:8088/oaiprovider/?verb=Identify

Workflow

Um Daten ins Repository einzuspielen nutzen wir eine bei uns entwickelte Java-Software, die ordnerweise Dateien auf Validität testet, Checksummen erstellt bzw. mit vorhandenen vergleicht und PIDs vergibt.

Korpora

- Die derzeit im Repositorium verfügbaren historischen Textkorpora stammen überwiegend aus dem DFG-Projekt „**D**eutsches **T**ext**a**rchiv“ (DTA, http://www.deutschestextarchiv.de) der BBAW.
- Die Volltexte des DTA sind anhand des auf XML/TEI P5-basierenden DTA-Basisformats annotiert, welches als Best-Practice Modell für geschriebene Korpora in CLARIN-D fungiert.
- **DTA-B**asis**f**ormat (DTABf): http://www.deutschestextarchiv.de/doku/basisformat.

Korpora: DTA und Dingler

Derzeit sind 1300 im Rahmen des DTA digitalisierten Werke im Repositorium verzeichnet, etwa 700 weitere sind in Arbeit. Seit kurzem ist auch der vollständige ‚Dingler', also sämtliche 370 Bände des von Johann Gottfried Dingler begründeten Polytechnischen Journals (Erscheinungszeitraum 1820–1831) im Repositorium zu finden. Das DTA bietet darüber hinaus unter http://www.deutschestextarchiv.de/download/ die Texte seines Kernkorpus sowie des bislang veröffentlichten Teils des Ergänzungskorpus zum Download an. Mit einem Zeitstempel versehen, können dort die gesamten Textdaten oder nach Genre (Belletristik – Gebrauchsliteratur – Wissenschaft) bzw. Erscheinungszeitraum zusammengestellte Teilkorpora als ZIP-Dateien heruntergeladen werden.
Weitere Informationen unter http://www.polytechnischesjournal.de/.

Korpora: Dingler und Wendekorpus

Die Volltexte aus dem ‚Dingler'-Projekt wurden ebenfalls entsprechend des DTABf annotiert und in die Korpusinfrastruktur des DTA integriert. Seit kurzem ist auch das ‚Berliner Wendekorpus' aus Transkripten von narrativen Interviews, die zwischen 1992 und 1996 mit Ost- und Westberlinern über deren persönliche ‚Wende'-Erfahrungen geführt wurden, im Repositorium verfügbar.

Lexikalische Ressourcen

Bei den im BBAW-Repositorium verwalteten lexikalischen Ressourcen handelt es sich um Stichwortlisten und Frequenzinformationen aus dem Bestand des „Digitalen Wörterbuchs der Deutschen Sprache“ (DWDS, http://www.dwds.de). Dazu gehören unter anderem

- die gegenwartssprachliche Wörterbuchkomponente des DWDS,
- das Etymologische Wörterbuch des Deutschen nach Wolfgang Pfeifer und
- die Erstbearbeitung des von Jacob Grimm und Wilhelm Grimm begründeten Deutschen Wörterbuchs ([1]DWB, 1854–1960).

Lexikalische Ressourcen

CLARIN-D

Über das Repositorium kann beispielsweise auf Stichwortlisten, angereichert mit grammatischen Informationen, aus den verschiedenen Wörterbüchern des DWDS zurückgegriffen werden. Hinzu kommt eine lexikalische Datenbank, die im dlexDB-Projekt (http://www.dlexdb.de/) erarbeitet wurde. Diese enthält Informationen über die Auftretenshäufigkeit von Wörtern und deren Kategorien, von Wortsequenzen sowie von sublexikalischen Einheiten wie z.B. Silben für den Einsatz in der psychologischen und linguistischen Forschung.

Externe Daten

Es ist auch für externe, d. h. nicht an der BBAW arbeitende Wissenschaftler möglich, Daten zur Langzeitarchivierung im Repositorium zu deponieren. Die Voraussetzungen dafür sind

- die inhaltliche Übereinstimmung der Daten mit den primären Aufgaben des BBAW-Repositoriums
- eine Konvertierung der Daten in die im Repositorium verwendeten Formate
- die Urheberrechte müssen für eine Aufnahme in das Repositorium geklärt sein.