

BAS Repository

Uwe Reichel
Institute of Phonetics and Speech Processing
University of Munich

31. März 2014



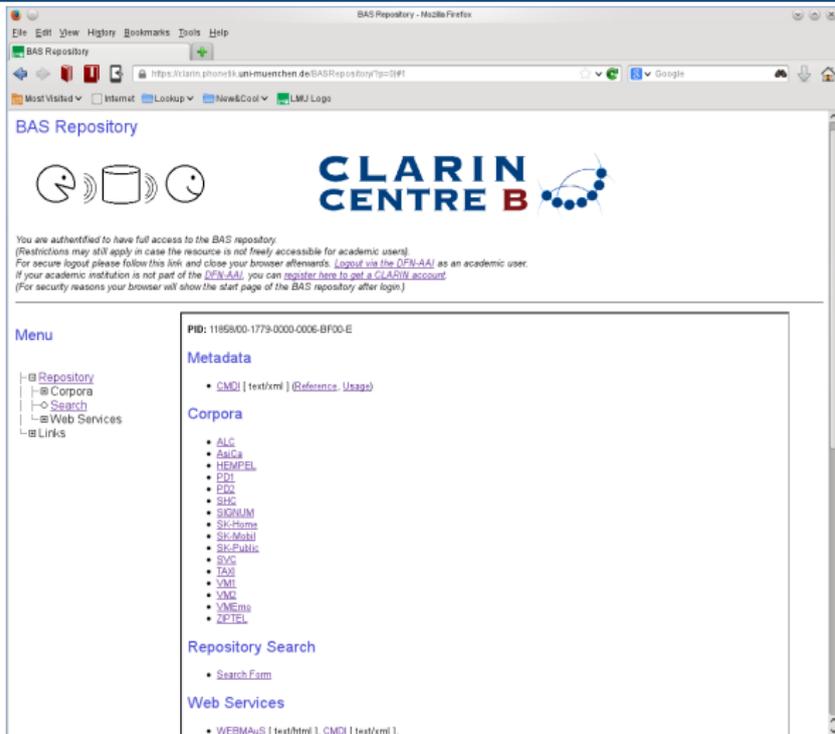
Inhalt

- **Was ist ein Daten-Repository?**
- **Wie kommen die Daten ins Repository?**
- **Wie kann man auf die Daten zugreifen?**

Was ist ein Daten-Repository?

- **Persistente** Speicherung von **Primärdaten** und **Metadaten**
- Daten werden **unverändert und dauerhaft** gespeichert.
- Studien sind replizierbar, Dokumentationen auffindbar, usw.
- Die Daten sind **dauerhaft zugänglich**, selbst wenn sich die URLs ändern sollten.
- Die Daten sind **geschützt**, d.h. nur über **Authentifizierung** zu erreichen.

BAS-Repository



BAS Repository - Mozilla Firefox

https://clarin.phonetik.uni-muenchen.de/BASRepository?ip=0#1

BAS Repository



You are authenticated to have full access to the BAS repository.
(Restrictions may still apply in case the resource is not freely accessible for academic users)
 For secure logout please follow this link and close your browser afterwards: [Logout via the OEN-AAI](#) as an academic user.
 If your academic institution is not part of the [OEN-AAI](#), you can [register here](#) to get a CLARIN account.
 (For security reasons your browser will show the start page of the BAS repository after login.)

Menu

- Repository
- Corpora
- Search
- Web Services
- Links

PID: 1185800-1779-0000-0006-BF00-E

Metadata

- [CMDI](#) [text/xml] [[Reference](#), [Usage](#)]

Corpora

- [SLC](#)
- [AsiCa](#)
- [HEMPFL](#)
- [PDI](#)
- [EDJ](#)
- [SHC](#)
- [SIGNUM](#)
- [SK-Home](#)
- [SK-Media](#)
- [SK-Public](#)
- [SVC](#)
- [TAXI](#)
- [VMI](#)
- [VMQ](#)
- [VMEms](#)
- [ZPTEL](#)

Repository Search

- [Search Form](#)

Web Services

- [WERMAoS](#) [text/html] [CMDI](#) [text/xml]



BAS-Repository

- 19 multimodale Sprachkorpora
- **öffentlicher Bereich**
 - Startseiten der Corpora und Sessions
 - Metadaten (CMDI-Format): 13 GByte
- **Geschützter Bereich**
 - Primärdaten (Signale, Annotationen): 2.5 TByte
- **Suchmaske**
- **OAI-PMH-Schnittstelle** zur Ausgabe der Metadaten
- **FCS-Schnittstelle** für *Federated Content Search*

Wie kommt ein Corpus ins Repository?

Was wird benötigt?

- Primärdaten
- ein CMDI-File zur Beschreibung des Corpus
- CMDI-Files zur Beschreibung der Aufnahme-Sessions

Wie kommt ein Corpus ins Repository?

Schritte

- 1 Metadaten werden validiert.
- 2 Metadaten werden in den öffentlichen Bereich kopiert und aktualisiert.
- 3 Primärdaten werden in den geschützten Bereich kopiert.
- 4 Für jedes File wird eine **Prüfsumme (Checksum)** ermittelt.
- 5 Für das Corpus und jede Session wird ein **Persistent Identifier (PID)** beantragt.
- 6 Suchdatenbank und Suchmaske werden re-initialisiert.
- 7 OAI-PMH-Schnittstelle wird reinitialisiert.

Persistente Archivierung

Prinzipien

- **Konsistenz:** Primärdaten dürfen im Repository **nicht mehr verändert** werden. Nur das Hinzufügen neuer Versionen ist erlaubt.
- **Persistenz:** Die Daten müssen **dauerhaft erhalten bleiben und erreichbar sein**.

Konsistenz mittels Prüfsumme

- Bitfolge einer Datei → eindeutiger Zahlen- und Buchstabencode
- Wenn diese Datei im Repository verändert wird, dann stimmt die neue Prüfsumme nicht mehr mit der gespeicherten überein.

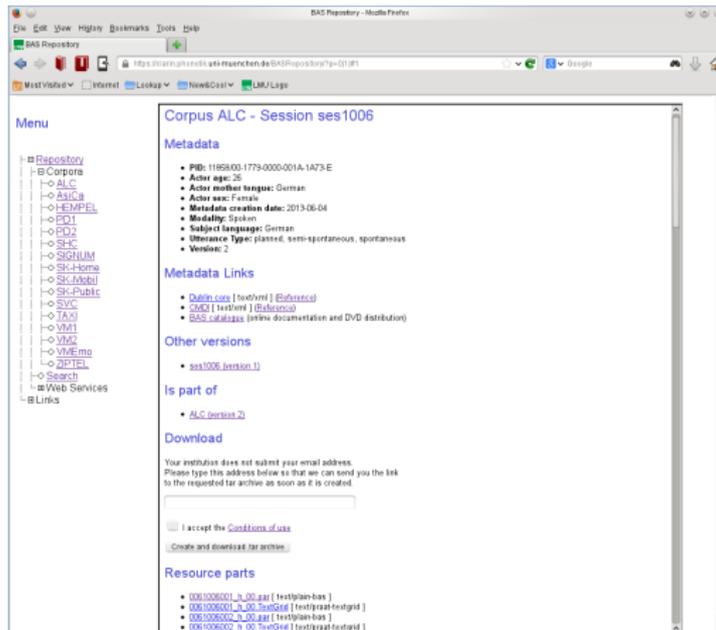
Persistente Archivierung

Persistenz: Persistent Identifier (PID)

- zur **eindeutigen Identifizierung** von Daten: jeder PID wird nur einmal vergeben und zeigt auf genau eine URL
- zum **dauerhaften Auffinden** der Daten: URLs können sich ändern, der PID bleibt unverändert
- zentrale Vergabe und Speicherung von PIDs und deren Verbindung zu URLs, z.B. *European Persistent Identifier Consortium (EPIC)*
- **Handle-System:** 11858/00-1779-0000-001A-1C5D-1
- Auflösung des PID zur Startseite:
<http://hdl.handle.net/11858/00-1779-0000-001A-1C5D-1>
- Auflösung des PID zu den Metadaten:
<http://hdl.handle.net/11858/00-1779-0000-001A-1C5D-1@format=cmdi>

Wie kann man auf die Daten zugreifen?

Landing Page: über die URL oder den PID zu erreichen.



The screenshot shows a web browser window titled "BAS Repository - Mozilla Firefox". The address bar contains the URL: <https://www.phil.uni-muenchen.de/BASRepository/?p=01181>. The page content is as follows:

Menu

- Repository
 - Corpora
 - ALC
 - AsiCa
 - HEMPEL
 - PD1
 - PD2
 - SHC
 - SIGNUM
 - SK-Home
 - SK-Mobil
 - SK-Public
 - SVC
 - TASJ
 - VMT
 - VMD
 - Vivemo
 - ZITEL
 - Search
 - Web Services
 - Links

Corpus ALC - Session ses1006

Metadata

- PID: 1192800-1779-0000-001A-1A73-E
- Actor age: 26
- Actor mother language: German
- Actor sex: Female
- Metadata creation date: 2013-06-04
- Modality: Spoken
- Subject language: German
- Utterance type: planned, semi-spontaneous, spontaneous
- Version: 2

Metadata Links

- Dublin core | text/xml | [Rdf/resource](#)
- SMD | text/html | [Rdf/resource](#)
- BSB catalog | [inflow documentation and DVD distribution](#)

Other versions

- ses1006 [version 1](#)

Is part of

- ALC [version 2](#)

Download

Your institution does not submit your email address. Please type this address below so that we can send you the link to the requested far archive as soon as it is created.

I accept the [Conditions of Use](#)

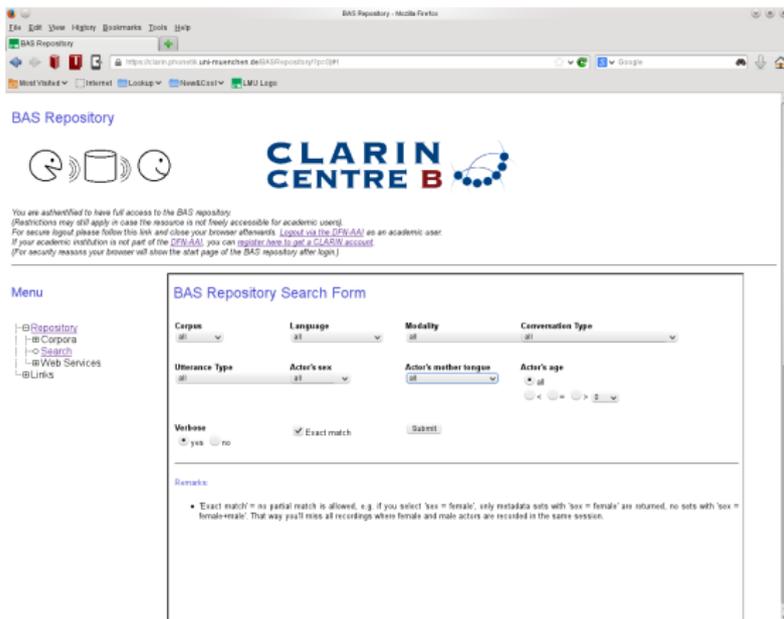
[Create and download far archive](#)

Resource parts

- 0061006001_00_001 | text/plain-bas |
- 0061006001_00_002 | text/plain-boolgrid |
- 0061006001_00_003 | text/plain-bas |
- 0061006001_00_004 | text/plain-boolgrid |

Wie kann man auf die Daten zugreifen?

Suchmaske



The screenshot shows a web browser window titled "BAS Repository - Mozilla Firefox". The address bar shows the URL "https://clarin.phonetik.uni-muenchen.de/BASRepository/". The page content includes the "BAS Repository" logo and the "CLARIN CENTRE B" logo. Below the logos, there is a notice about user authentication and a "Menu" on the left side. The main content area is titled "BAS Repository Search Form" and contains several search filters:

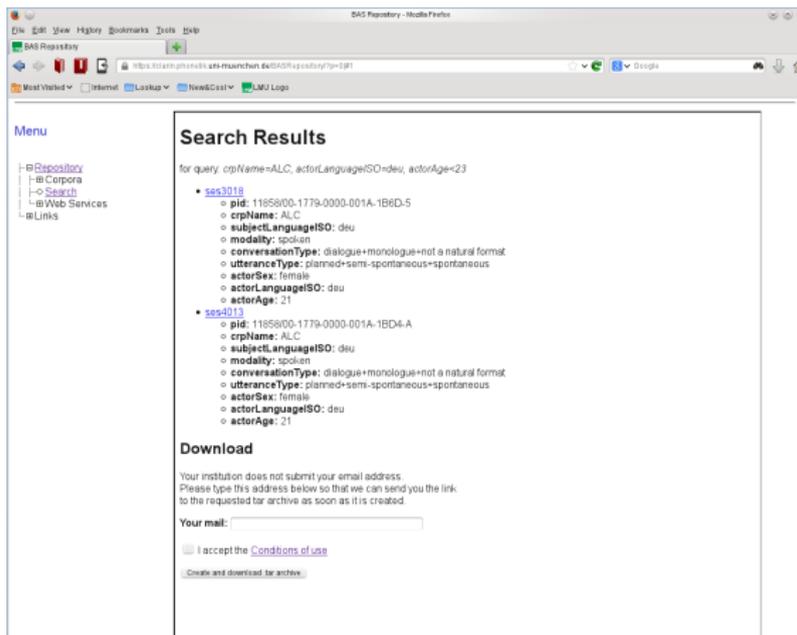
- Corpus:** dropdown menu set to "all"
- Language:** dropdown menu set to "all"
- Modality:** dropdown menu set to "all"
- Conversation Type:** dropdown menu set to "all"
- Utterance Type:** dropdown menu set to "all"
- Actor's sex:** dropdown menu set to "all"
- Actor's mother tongue:** dropdown menu set to "all"
- Actor's age:** range selector set to "all"
- Verbose:** radio buttons for "yes" and "no", with "Exact match" checked.
- Submit:** button

Below the search form, there is a "Remarks" section with a bullet point:

- "Exact match" = no partial match is allowed, e.g. if you select "sex = female", only metadata sets with "sex = female" are returned, no sets with "sex = female-male". That way you'll filter all recordings where female- and male-actors are recorded in the same session.

Wie kann man auf die Daten zugreifen?

Suchmaske



Search Results

for query: crpName=ALC, actorLanguageISO=deu, actorAge=23

- ses3018**

 - pid: 11858/00-1779-0000-001A-1B6D-5
 - crpName: ALC
 - subjectLanguageISO: deu
 - modality: spoken
 - conversationType: dialogue+monologue+not a natural format
 - utteranceType: planned+semi-spontaneous+spontaneous
 - actorSex: female
 - actorLanguageISO: deu
 - actorAge: 21
- ses4013**

 - pid: 11858/00-1779-0000-001A-1BD4-A
 - crpName: ALC
 - subjectLanguageISO: deu
 - modality: spoken
 - conversationType: dialogue+monologue+not a natural format
 - utteranceType: planned+semi-spontaneous+spontaneous
 - actorSex: female
 - actorLanguageISO: deu
 - actorAge: 21

Download

Your institution does not submit your email address.
Please type this address below so that we can send you the link
to the requested file as soon as it is created.

Your mail:

I accept the [Conditions of use](#)

[Create and download file archive](#)

Wie kann man auf die Daten zugreifen?

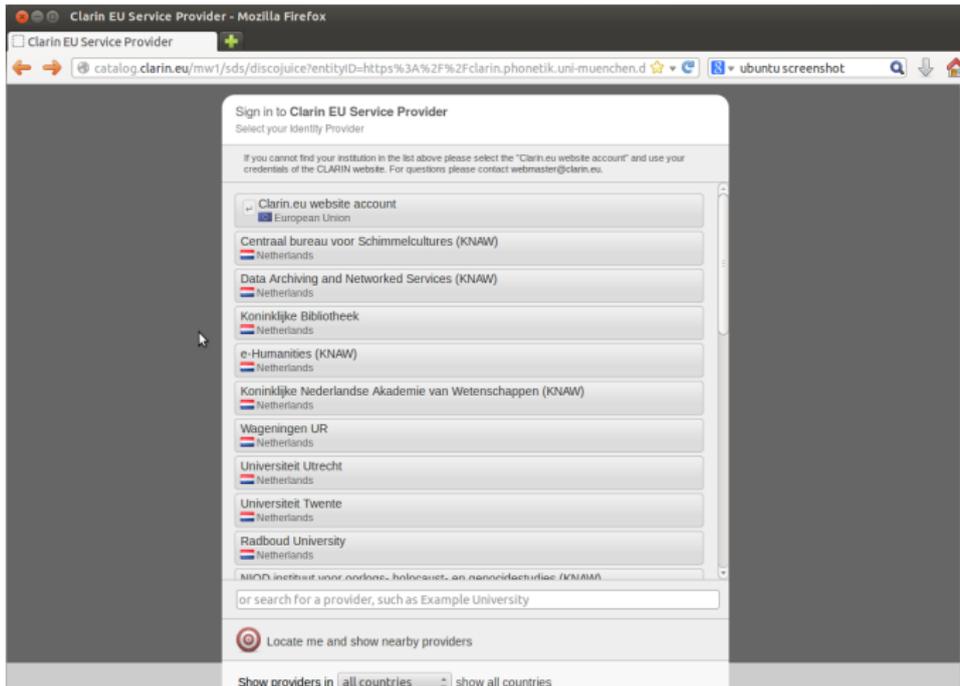
OAI-PMH-Schnittstelle

- Ausgabe der Metadaten gemäß dem **O**pen **A**rchive **I**nitiative **P**rotocoll for **M**etadate **H**arvesting
- für **Harvester**, die das WWW nach Metadaten absuchen, beispielsweise für das **Virtual Language Observatory**

<http://www.phonetik.uni-muenchen.de/cgi-bin/BASRepository/oaipmh/oai.pl>

- **Wer ist die Schnittstelle? Welche Metadatenformate werden ausgeliefert?** `?verb=Identify` `?verb=ListMetadataFormats`
- **Zeige alle Identifier zum Corpus ALC!**
`?verb=ListIdentifiers&metadataPrefix=cmdi&set=Corpora:ALC`
- **Zeige den Inhalt zu einem bestimmten Identifier!**
`?verb=GetRecord&identifier=oai:BAS.repo:Corpora/ALC/ses1006&metadataPrefix=cmdi`

Authentifizierung



Clarin EU Service Provider - Mozilla Firefox

Clarin EU Service Provider

catalog.clarin.eu/mw1/sds/discojuice?entityID=https%3A%2F%2Fclarin.phonetik.uni-muenchen.d

ubuntu screenshot

Sign in to Clarin EU Service Provider

Select your Identity Provider

If you cannot find your institution in the list above please select the "Clarin.eu website account" and use your credentials of the CLARIN website. For questions please contact webmaster@clarin.eu.

- Clarin.eu website account
 - European Union
- Centraal bureau voor Schimmelcultures (KNAW)
 - Netherlands
- Data Archiving and Networked Services (KNAW)
 - Netherlands
- Koninklijke Bibliotheek
 - Netherlands
- e-Humanities (KNAW)
 - Netherlands
- Koninklijke Nederlandse Akademie van Wetenschappen (KNAW)
 - Netherlands
- Wageningen UR
 - Netherlands
- Universiteit Utrecht
 - Netherlands
- Universiteit Twente
 - Netherlands
- Radboud University
 - Netherlands
- MIND institute voor online, holistische, en narratieve taal (KNAW)

or search for a provider, such as Example University

Locate me and show nearby providers

Show providers in show all countries

Authentifizierung

- Anmeldung über die jeweilige akademische Einrichtung.
- Die Einrichtung übermittelt dem Türwächter die Information zum *Entitlement*.
- Ist das Entitlement **academic**, dann wird der Zugang freigegeben.
- Alternativ können Zugänge selektiv auch über Mailadressen freigeschaltet werden.