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?**

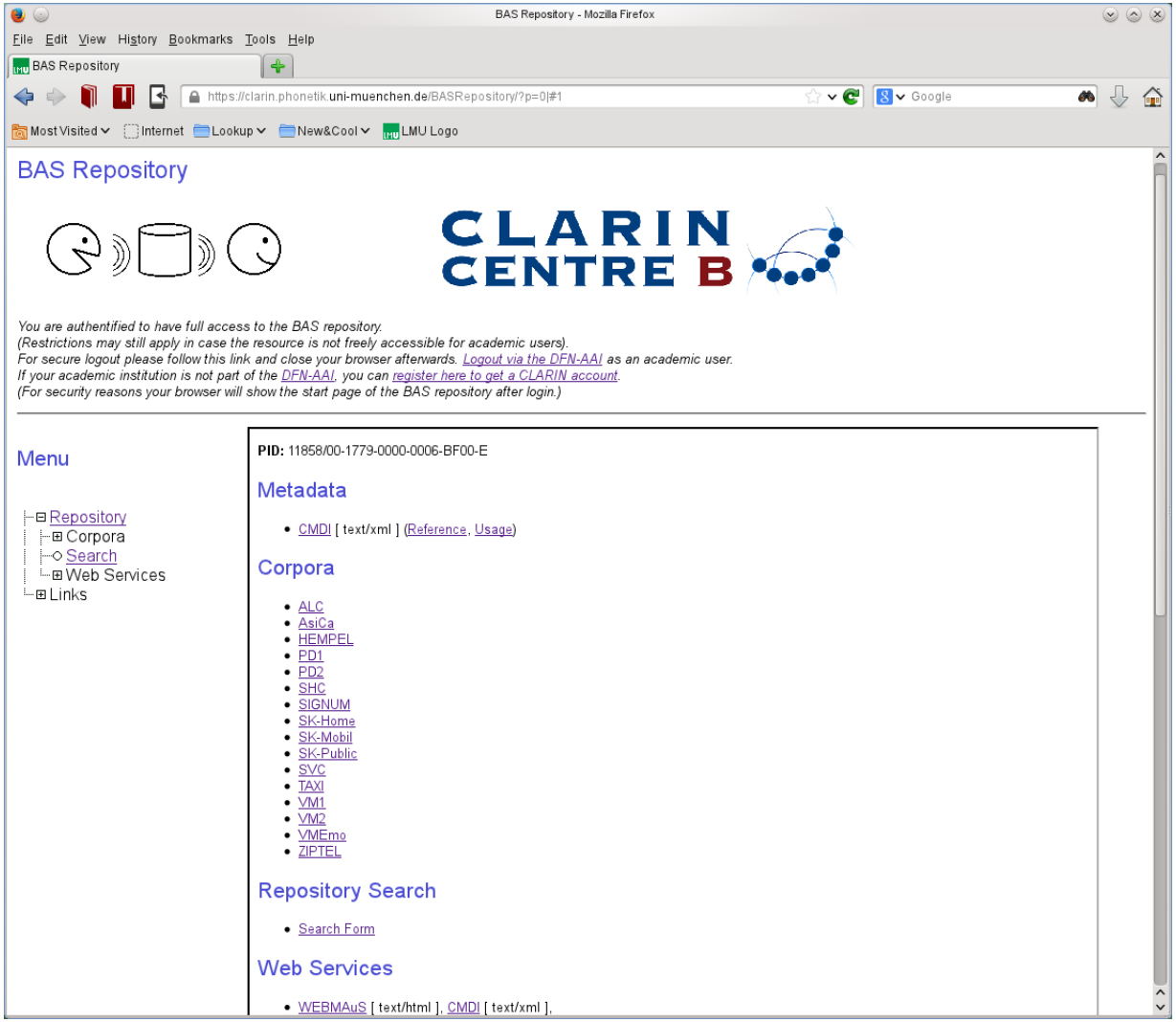
Was ist ein Daten-Repository?

- **Persistente** Speicherung von **Primärdaten** und **Metadaten**
- Daten werden **unverändert und dauerhaft** gespeichert.

$\longrightarrow$ 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

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

Persistente Archivierung

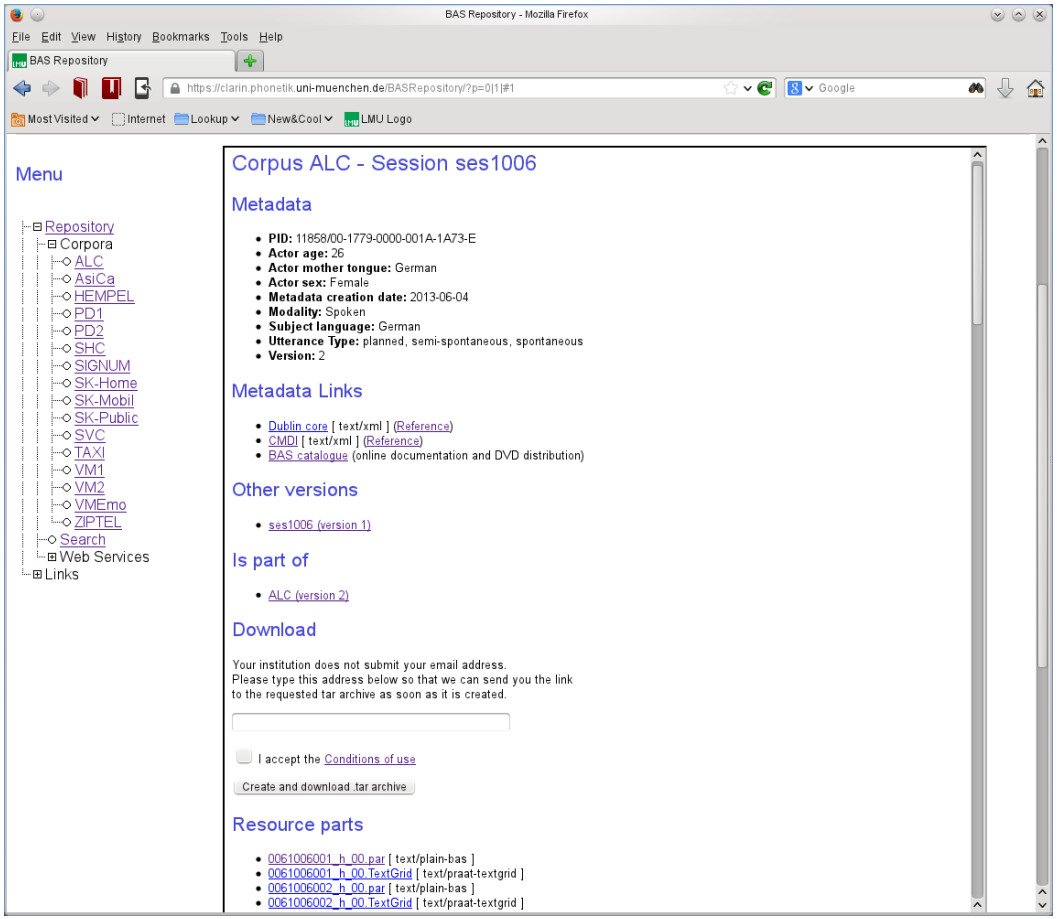
Persistenz: Persistent Identifier (PID)

- 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

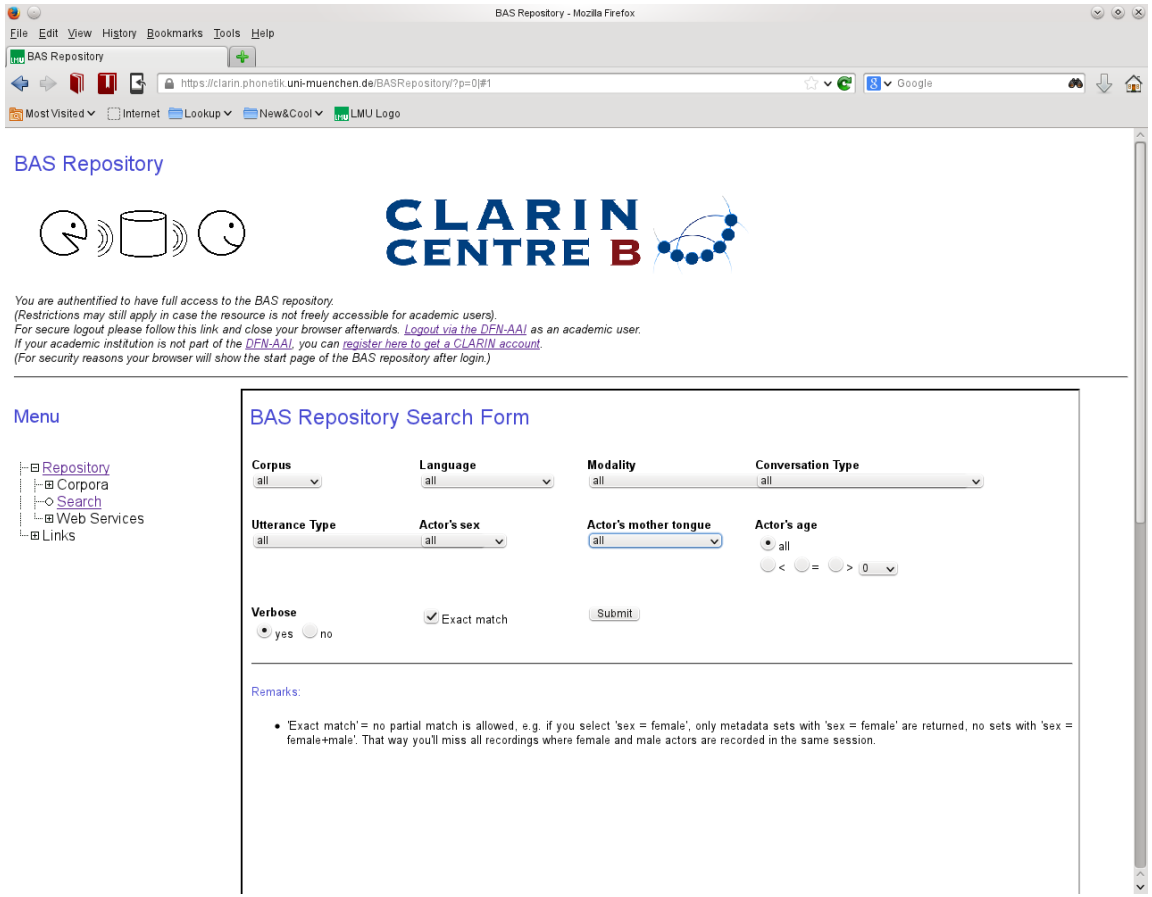
Wie kann man auf die Daten zugreifen?

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



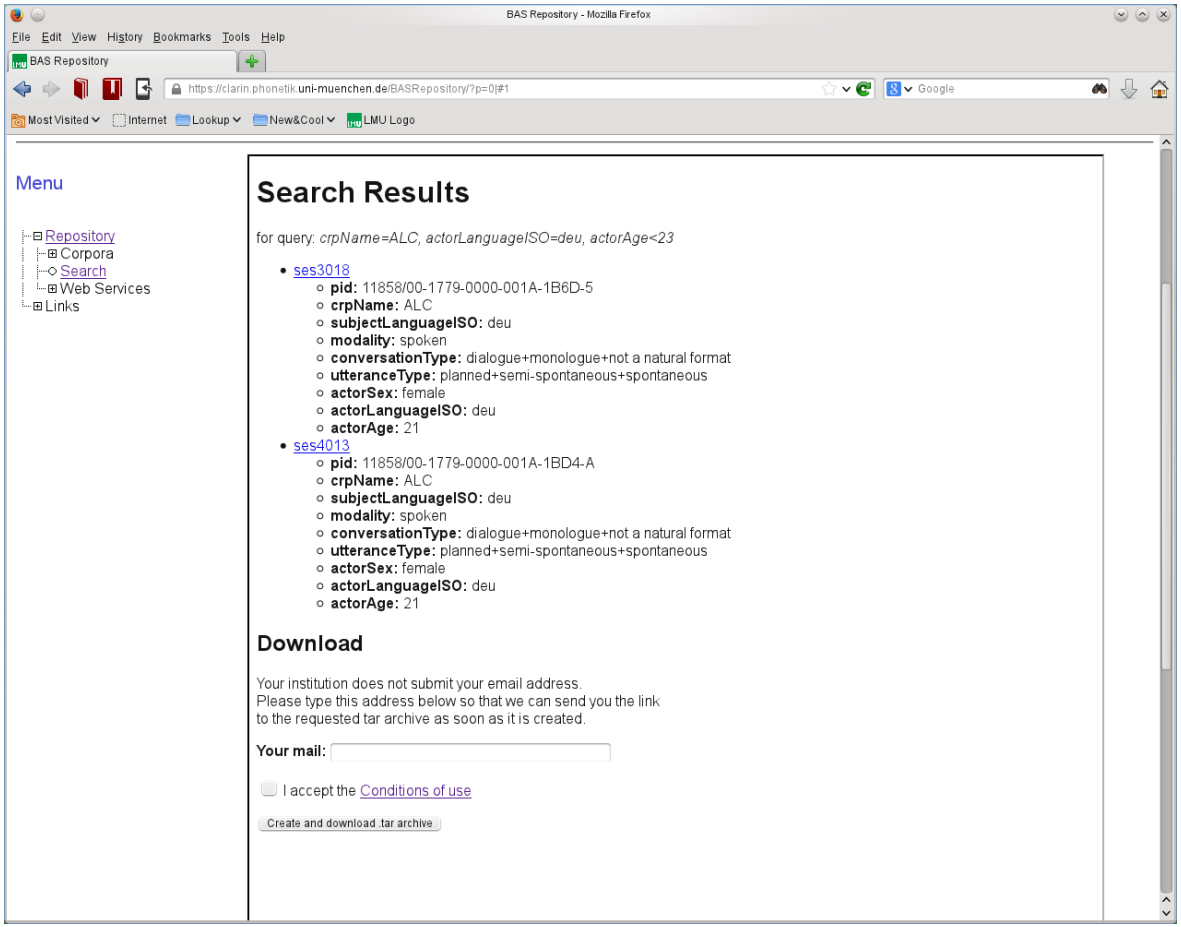
Wie kann man auf die Daten zugreifen?

Suchmaske



Wie kann man auf die Daten zugreifen?

Suchmaske



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**etadata **H**arvesting
- für **Harvester**, die das WWW nach Metadaten absuchen, beispielsweise für das **Virtual Language Observatory**

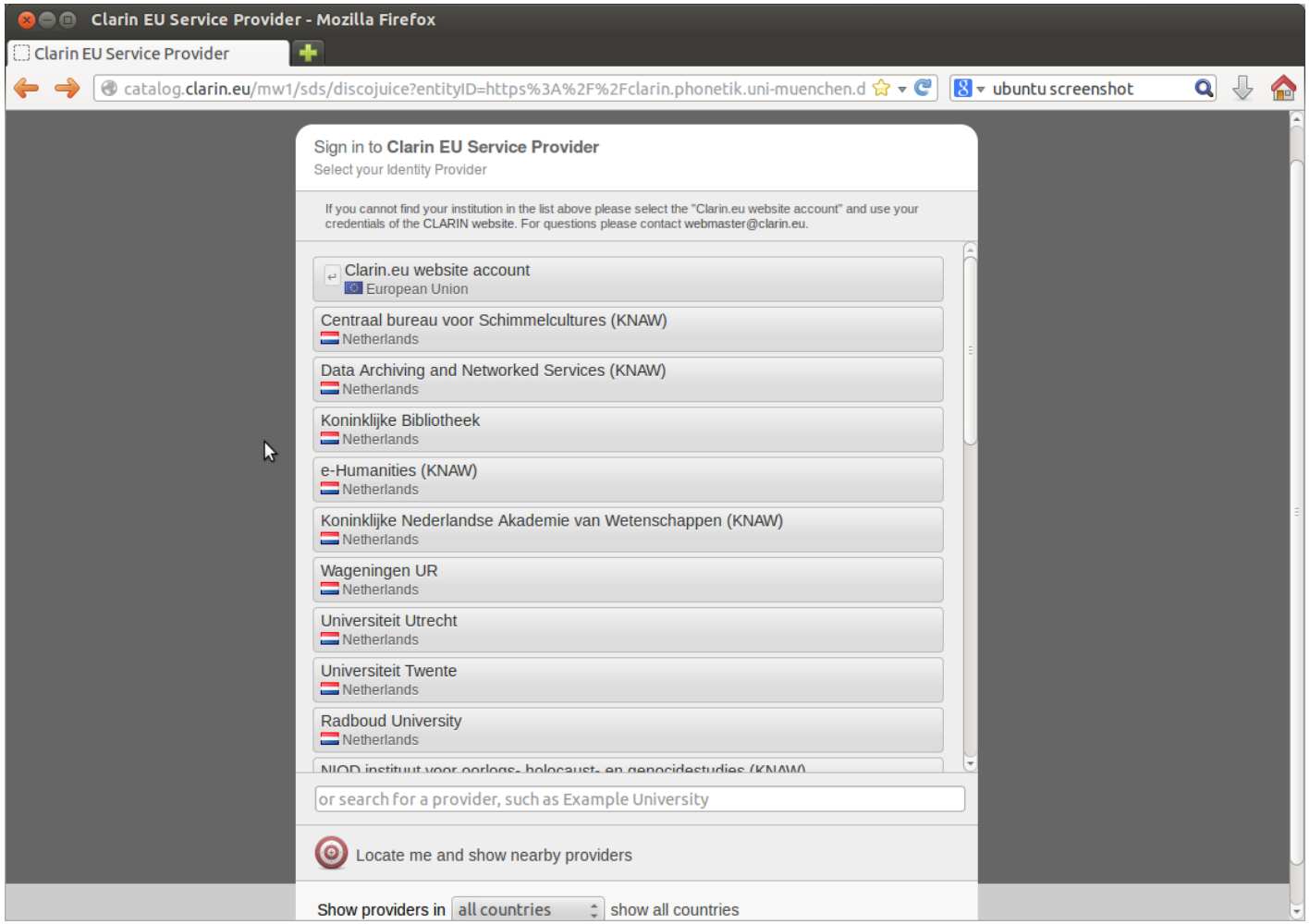
`http://www.phonetik.uni-muenchen.de/cgi-bin/BASRepository/oaipmh/oai.pl`

- **Wer ist die Schnittstelle?** `?verb=Identify` **Welche Metadatenformate werden ausgeliefert?** `?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



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.