Das BAS-Repository

Uwe D. Reichel
Institut für Phonetik und sprachverarbeitung
reichelu@phonetik.uni-muenchen.de

1 Aufbau

Das BAS-Repository ist über die folgende Webseite zu erreichen:

`https://clarin.phonetik.uni-muenchen.de/BASRepository/`

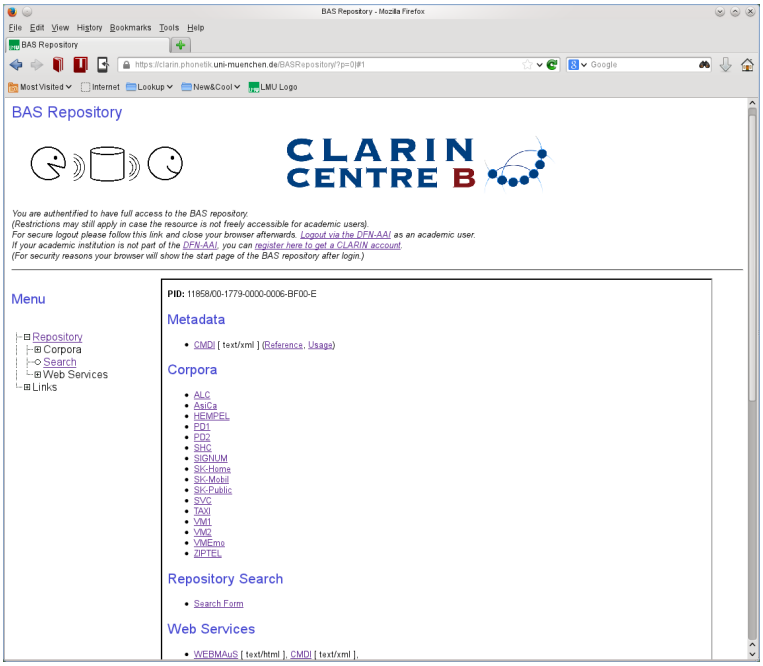

Abbildung 1: Startseite des Repositories.

Es umfasst zum gegenwärtigen Zeitpunk 16 multimodale Sprachkorpora mit einem Umfang von insgesamt 2.5 TByte sowie zugehörigen Metadaten im CMDI-Format im Umfang von 13 GByte.

Dem Repository liegt ein Filesystem zugrunde mit einem frei zugänglichen und einem geschützten Bereich. Frei zugänglich sind die Startseiten

sowie die Metadaten der Repository-Objekte. Im geschützen Bereich befinden sich die Resourcen, also die Signaldateien und Annotationen.

Die Objekte des BAS-Repositories sind Versionen von Korpora und Aufnahmesessions, wobei ein Korpus aus einer oder mehreren Sessions besteht. Jedes dieser Objekte wird durch ein eigenes CMDI-File beschrieben und ist über eine eigene Startseite zugänglich, die dynamisch aus dem CMDI-File erzeugt wird.

Die Metadaten können über eine OAI-PMH-Schnittstelle geharvestet werden:

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

2 Zugang

Jedem Repository-Objekt ist ein EPIC Handle Persistent Identifier (PID) zugeordnet, über den es dauerhaft zugänglich ist, wie beispielsweise die Startseite der Session 1006 des Korpus ALC:

`http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3`

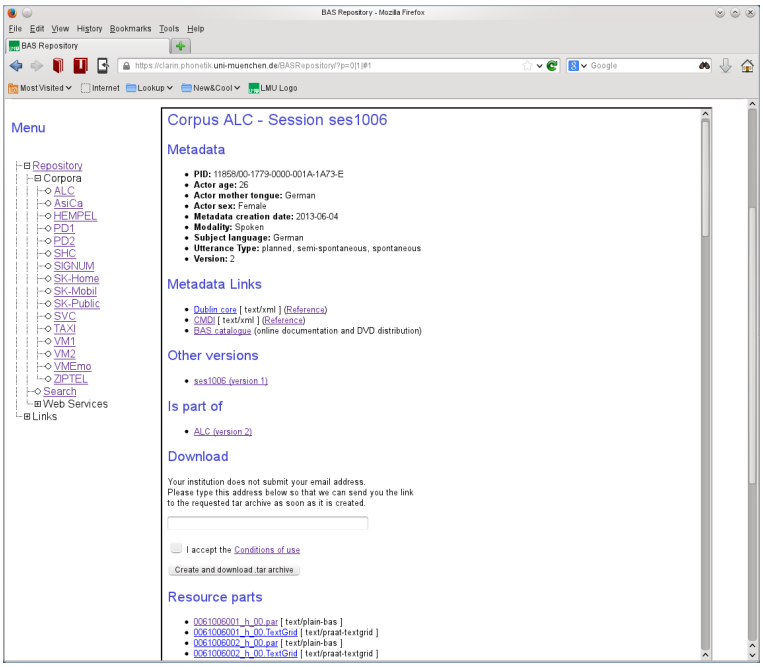

Abbildung 2: Landing Page einer Aufnahme-Session.

Neben einer Kurzbeschreibung findet sich hier auch ein Link zu den kompletten Metadaten. Diese lassen sich für eine automatisierte Verarbeitung

auch über zwei direkte Wege abrufen: Zum einen ist die Verwendung eines Part Identifiers @format=cmdi möglich:

`http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3@format=cmdi`

Über diesen Weg können die Metadaten auch im kompakteren Dublin-Core-Format mittels @format=dc angezeigt werden. Ein weiterer direkter Zugriff auf Metadaten wird mittels Content negotiation ermöglicht, indem client-seitig der Accept Header auf applicationx-cmdi+xml gesetzt wird.

Im Gegensatz zu den frei zugänglichen Startseiten und Metadaten liegen die primären Ressourcen in einem Shibboleth-geschützten Bereich, der erst nach entsprechender Autorisierung zugänglich ist. Den Nutzern werden drei Möglichkeiten geboten, sich zu authentifizieren. Alle akademischen Nutzer, deren Institution Teil des DFN-AAI-Netzwerks ist, können sich über die DFN-AAI authentifizieren. Akademische Nutzer anderer Institute beantragen eine Clarin-Kennung und authentifizieren sich über diese. Nicht-akademische Kunden können sich nach Zulegung einer Clarin-Kennung selektiv Rechte für Einzelkorpora erwerben.

Nach erfolgter Autorisierung werden die Links zu den Ressourcen auf der Startseite angezeigt sowie ein direkter Zugriff über den Part-Identifier @partId ermöglicht, dessen Werte den Resource-Proxy-Ids in den CMDI-Files entsprechen. Beispiel:

`http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3@partId=m_0000000001`

Zudem erhält der autorisierte Nutzer die Möglichkeit, das entsprechende Repository-Objekt als komprimiertes tar-Archiv herunterzuladen.

3 Suchmaske

Über eine Suchmaske kann der Nutzer korpusübergreifend Aufnahmesessions für spezielle Forschungsfragen zusammenstellen wie beispielsweise im Hinblick auf Geschlecht oder Muttersprache der Sprecher.

Nach erfolgter Autorisierung lässt sich die gewünschte Auswahl als komprimiertes tar-Archiv herunterzuladen.

4 Aufnahme neuer Daten

Die Aufnahme eines neuen Korpus in das BAS-Repository verläuft vollautomatisch in folgenden Schritten:

1. Die CMDI-Files werden validiert, eingelesen und mit dem Inhalt einer Repository-Content-Tabelle abgeglichen, um festzustellen, ob es sich um neue Daten oder ein Update bereits gespeicherter Daten handelt.

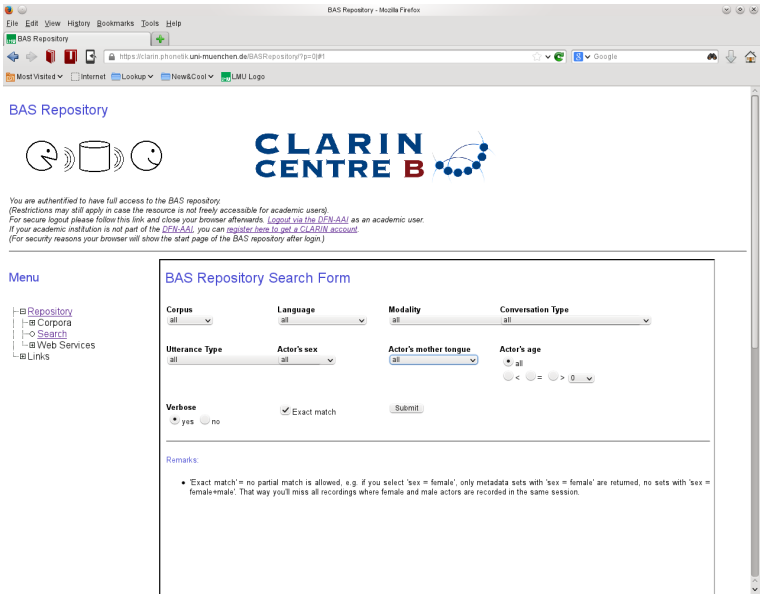

Abbildung 3: Suchmaske des Repositories.

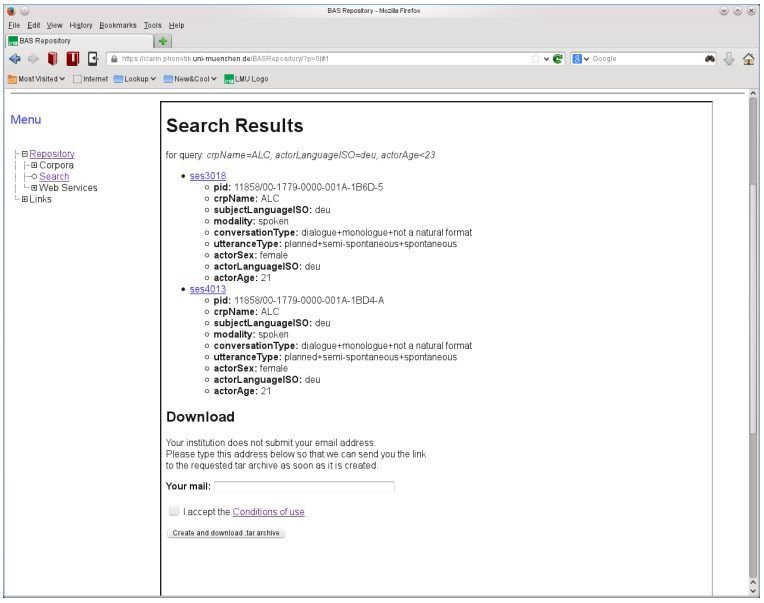

Abbildung 4: Ergebnis einer Suche im Repository.

2. Für alle neuen, beziehungsweise aktualisierten Sessions sowie das Corpus werden Persistent Identifier beantragt. Jede Version eines Corpus und einer Session erhält somit einen eigenenen Identifier.

3. CMDI-Filen werden in den frei zugänglichen Bereich kopiert und ange-

passt. Die Resourcen werden in den geschützten Bereich kopiert. Für regelmäßige Konsistenzprüfungen und für die Versionierung werden Checksums ermittelt.

4. Abschließend erfolgt eine Aktualisierung der Suchdatenbank sowie der an der OAI-PMH-Schnittstelle gespeicherten Daten.

Auch BAS-externe Korpora können auf diese Weise gehostet werden. Alle Daten werden regelmäßig durch das Leibnitz Rechenzentrum in Garching durch Backups gesichert.

5 Software

Für das BAS-Repository wurde eine proprietäre Softwarelösung in Perl und PHP entwickelt. Voraussetzungen sind ein CGI- und PHP-fähiger Server, SQLite als Backend der Suchmaske, sowie frei erhältliche Tools zur XML-Validierung, Metadaten-Transformation und Checksum-Ermittlung. Für die OAI-PMH-Schnittstelle wurde der frei erhältliche OAI-PMH2 XMLFile Datenprovider angepasst.