Bayerisches Archiv für Sprachsignale (BAS) - Langzeitarchivierung von Interview-Sammlungen

Informationen für Sammlungsinhaber (Florian Schiel - 13.03.2021)

Synopsis

Diese Dokument dient zur Information für Inhaber von Oral History-Sammlungen, die das Portal „Oral-History.Digital“ der Freien Universität Berlin nutzen wollen (im Folgenden „Sammlungsinhaber“ genannt). Es beschreibt die im Rahmen des Projektes „Oral-History.Digital“ angebotene Langzeitarchivierung von Interview-Sammlungen am Bayerischen Archiv für Sprachsignale an der Ludwig-Maximilians-Universität München.
Die Langzeitarchivierung im BAS CLARIN Repository bietet dem Sammlungsinhaber eine dauerhafte Datensicherung nach höchsten Sicherheitsstandards mit langjähriger Erfahrung und vielfältigen Services.

1. Allgemeines zum BAS

Das Bayerische Archiv für Sprachsignale (BAS) wurde im Januar 1995 als öffentliche Einrichtung an der Ludwig-Maximilians-Universität (LMU) München gegründet. Das BAS befindet sich derzeit am Institut für Phonetik und Sprachverarbeitung (https://www.phonetik.uni-muenchen.de/).
Zentrale Aufgabe des BAS ist es, digitale Sprachressourcen basierend auf gesprochenem Deutsch in strukturierter Form zu archivieren und sowohl der Forschungsgemeinschaft als auch der Sprachtechnologie verfügbar zu machen.
Seit 1995 sammelt, produziert, standardisiert, pflegt und distribuiert das BAS digitale Sprachressourcen für gesprochenes Deutsch. Seit 2009 ist das BAS Mitglied im CLARIN-Verbund zum Aufbau einer europäischen Infrastruktur für Sprachdaten und Sprachtechnologie (http://www.clarin.eu). Seit 2016 ist das BAS zertifiziertes CLARIN B Centre; das BAS wurde außerdem mit dem "Core Trust Seal" zertifiziert (https://www.coretrustseal.org/).
Derzeit (März 2021) pflegt in seinem Repositorium das BAS 52 Sammlungen mit ca. 28.000 Aufnahmen; davon sind 3 Sammlungen öffentlich, 46 frei zugänglich für Akademiker und 6 zugänglich nach Anmeldung beim Sammlungsinhaber. Die Startseite des BAS CLARIN Repositories ist unter der dauerhaften Adresse (Persistent Identifier) http://hdl.handle.net/11022/1009-0000-0001-231F-6 erreichbar.
Am BAS werden außerdem eigene Verfahren zur automatischen Aufnahme, Verarbeitung, Etikettierung und Segmentierung von Sprachdaten entwickelt, die entweder als public domain Software oder als Webservices zur Verfügung gestellt werden.
Das BAS wird von der LMU München mit zwei Dauerstellen und Infrastruktur finanziert; darüber hinaus bestehen öffentliche Förderungen (BMBF, BMWi, DFG). Zudem verfügt das BAS über Einkünfte aus dem Verkauf von Nutzerlizenzen an nicht-akademische Nutzer.

2. BAS und Oral-History.Digital

Im Rahmen des DFG-geförderten Projektes „Oral-History.Digital“ bietet das BAS Sammlungsinhabern eine Langzeitarchivierung (LZA) ihrer Interview-Sammlungen im BAS CLARIN Repository an. Das bedeutet, dass die Interview-Sammlungen dauerhaft gesichert und gepflegt werden und bei Zustimmung durch die Sammlungsinhaber Dritten zur Verfügung gestellt werden können. Das betrifft vor allem die hochwertigen Original- oder Archivversionen der Audio- oder Videoaufnahmen.
Im Gegensatz dazu stellt die von der Freien Universität Berlin betriebene Online-Plattform „Oral-History.Digital“ internetfähige, komprimierte Mediendateien bereit, die in einer nutzerfreundlichen Redaktions- und Rechercheumgebung komfortabel erschlossen und sammlungsübergreifend recherchiert werden können.
Im Folgenden beschreiben wir kurz die Vorteile einer solchen LZA, gefolgt von ausführlicheren Beschreibungen der Voraussetzungen für eine LZA, des Kostenmodells, der Zugangskontrolle (Nutzungslizensierung) und grundsätzliche Fragen der Beständigkeit des BAS.

3. Vorteile der LZA einer Interview-Sammlung am BAS

Das BAS Repository orientiert sich in all seinen Prozessen an den FAIR-Prinzipien für Forschungsdaten: *Findability, Accessibility, Interoperability, Reusability*:

- *Findability*: Die im BAS Repository archivierten Interviews sind **in übergeordneten Verzeichnissen auffindbar**. Metadaten des BAS sind standardisiert (DC, OLAC, CMDI) und werden über eine automatisierte Schnittstelle (OAI-PMH) an *Science Indices* verbreitet; Dadurch werden Ihre Daten sofort sichtbar, z.B. im Virtual Language Observatory (https://vlo.clarin.eu/), im Reuter Science Data Index, in der Open Language Archives Community (http://www.language-archives.org/). Die Metadaten des BAS sind grundsätzlich öffentlich und dürfen daher keine personenbezogenen Daten enthalten.
- *Findability*: Die im BAS Repository archivierten Interviews sind **an einer dauerhaften Webadresse auffindbar**. Für jede Sammlung (und jedes Interview innerhalb der Sammlung) werden *Persistent Identifiers* (PID) im *handle system* (https://www.handle.net/) beantragt und gepflegt. PIDs erlauben eine dauerhafte Referenzierung (z.B. in wissenschaftlichen Veröffentlichungen) eines archivierten Datensatzes. Wenn eine PID verwendet wird (z.B. die URL des BAS Repositories oben), kann man sicher sein, dass diese immer auf den richtigen Datensatz zeigt, auch wenn das Archiv selber inzwischen umgezogen ist. Hier ist eine Beispiel-PID einer Sammlung im BAS: http://hdl.handle.net/11022/1009-0000-0001-88E5-3
- *Accessibility*: Die im BAS Repository archivierten Interviews sind **vor unbefugtem Zugang geschützt**. Der Sammlungsinhaber entscheidet in drei Lizenzmodellen, wer auf seine Daten in der LZA zugreifen darf: „PUB“ ist für jedermann zugänglich, „ACA“ nur für Akademiker, die sich über das AAI Shibboleth-System als Angehörige einer Universität ausweisen können, oder „RES“ nur für den Sammlungsinhaber oder von ihm zugelassene Personen. Beim „RES“-Modell ist also langfristig sichergestellt, dass nur vom Sammlungsinhaber autorisierte Personen Zugang zu den geschützten Audio- oder Videodateien und Begleitmaterialien haben; nur wenige, anonymisierte Metadaten sind immer öffentlich zugänglich.
- *Accessibility*: Die im BAS Repository archivierten Interviews sind **dauerhaft gegen Verlust gesichert**. Daten werden professionell nach den Prinzipien des Open Archival Information System (OAIS) gesichert; dazu gehören tägliche Backups an zwei Standorten (Leibniz-Rechenzentrum München, Rechenzentrum Jülich). Nur autorisiertes IT-Personal des BAS hat direkten Zugriff auf die Daten. Die ursprünglich zur Verfügung gestellten Daten können bei Bedarf immer aus der LZA rekonstruiert werden.
- *Accessibility*: Die im BAS Repository archivierten Interviews sind für den Sammlungsinhaber **jederzeit wieder leicht zugänglich**. Die Archivierung ist ein Live-System, d.h. die Daten sind für den Sammlungsinhaber (und für von ihm autorisierte Nutzer) jederzeit abrufbar. Das ist wesentlich komfortabler als eine Rückholung bei der üblichen Sicherung auf LTO-Magnetbändern in anderen Rechenzentren.
- *Interoperability*: Die im BAS Repository archivierten Interviews sind **stets im aktuellen Medienformat abspielbar**. Daten werden regelmäßig auf ihre Integrität geprüft. Dabei wird auch geprüft, ob das Mediaformat weiterhin in aktuell verbreiteten Playern abspielbar ist; ggf. werden Daten auf neue Formate transkodiert.
- *Reusability*: Die im BAS Repository archivierten Interviews sind **auch für andere Disziplinen nutzbar**. Oral History-Interviews sind wertvolle Quellen auch für andere Disziplinen als die Geschichtswissenschaften; über das BAS Repository werden sie insbesondere für sprachwissenschaftliche Untersuchungen sichtbar.

4. Voraussetzungen für eine Nutzung des LZA für Sammlungsinhaber

(s. auch: https://www.phonetik.uni-muenchen.de/Bas/BasPolicyExternalResources_deu.pdf)

- Archiviert werden nur abgeschlossene (Teil-)Sammlungen von Interviews, bei denen keine Änderungen an den Audio- oder Mediendateien mehr erfolgen. Filmschnitt, Tonbearbeitung, Anonymisierung etc. müssen vor der LZA erfolgt sein, Korrekturen/Ergänzungen in Begleitdaten wie Transkript, Kurzbiografie etc. können dagegen nachgereicht werden.
- Die Erstellung der Sammlung wurde nach anerkannten ethischen Grundregeln und wissenschaftlichen Standards durchgeführt.
- Der Sammlungsinhaber kann nachweisen, dass er das Recht hat, die Interview- und Metadaten an das BAS zu übergeben; entweder liegen schriftliche Einverständniserklärungen der Interviewten vor, oder der Sammlungsinhaber erklärt gegenüber dem BAS, die notwendigen Rechte zu haben.
- Eine Sammlung muss Metadaten, pro Interview mindestens eine Mediendatei und mindestens eine Form der Verschriftung oder Annotation enthalten (letztere kann auch nachgereicht werden).
- Der Sammlungsinhaber muss die nicht-ausschließlichen Rechte zur Verwaltung, Pflege, Archivierung und evtl. zur Weitergabe an Dritte an das BAS übertragen; das geschieht in Form eines Überlassungsvertrags. (s. Beispiel in https://www.phonetik.uni-muenchen.de/Bas/BasTemplateContract.pdf)
- Eine Erhebung von Nutzungsgebühren durch das BAS oder durch den Sammlungsinhaber sind im Rahmen des Projekts Oral-History.Digital nicht vorgesehen. Wenn der Sammlungsinhaber einem Nutzer den Zugang zu seinen Daten gewährt, erfolgt dieser Zugang kostenlos.

5. Kosten

Das BAS bemüht sich, für die Mehrzahl der aufgenommenen Sammlungen die Kosten für Ingest, Dokumentation, Pflege und Hardware selber zu übernehmen. Für umfangreiche Datensätze fallen allerdings einmalige Gebühren an, welche die laufenden Kosten wenigstens teilweise abdecken sollen:

- Sammlungsumfang < 20GB : kostenfrei
- Sammlungsumfang 20GB - 1000GB : € 0,50 pro GB
- Sammlungsumfang > 1000GB : € 0,30 pro GB

Beispiel: Für einen insgesamten Datenumfang von 1500GB erhebt das BAS eine einmalige Gebühr von: €0 (0-20GB) + €490 (20-1000GB) + €150 (1000-1500GB) = €640

Diese Gebühren sind Nettopreise und enthalten bereits den üblichen 20% Overhead. Hierzu stellt das BAS dem Sammlungsinhaber nach Abschluss des Überlassungsvertrags eine Rechnung.

Kommt es bei der Aufnahme einer Sammlung zu einen personellen Mehraufwand für die Aufbereitung der Daten, z.B. weil Metadaten fehlerhaft sind oder Mediendateien transkodiert werden müssen, berechnen wir aktuell (2021) pro Arbeitsstunde € 35 (Spezialist) oder € 20 (manuelle Arbeiten).

6. Kontrolle durch den Sammlungsinhaber, Veränderungen, Löschung

Der Sammlungsinhaber hat jederzeit Zugriff auf seine Sammlung.

Der Sammlungsinhaber kann jeweils einmal pro Halbjahr Metadaten, Transkripte und andere Begleitdaten der Sammlung hinzufügen; neue Versionen der Mediendaten werden nur in Ausnahmefällen angenommen.

Der Sammlungsinhaber hat kein Recht auf Löschung der Daten, außer es handelt sich um zwingende Gründe, wie z.B. die Einhaltung von datenschutzrechtlichen Vorschriften oder von ethischen Grundsätzen. Löschungen sind kostenpflichtig und werden nach Aufwand in Rechnung gestellt.

Der Grund, warum Löschungen weitgehend vermieden werden sollten, ist, dass das BAS Persistent Identifiers für alle archivierten Mediendaten beantragt; diese können nicht mehr gelöscht werden,

d.h. das BAS ist verpflichtet, immer eine URL für jede PID bereitzustellen; nur dadurch wird das Grundprinzip von PIDs gewahrt.

7. Beständigkeit des BAS

Das BAS existiert in seiner Form seit 1995. Die LMU München als Host-Organisation ist daran interessiert, die Institution BAS zeitlich unlimitiert weiter zu betreiben. Aus rechtlichen Gründen gibt es eine Bestandsgarantie im Moment aber nur bis 2029.
Für den unwahrscheinlichen Fall, dass das BAS an seiner jetzigen Host-Organisation nicht weiter betrieben werden kann, besteht eine gegenseitige Bestandsgarantie zwischen den europäischen CLARIN Datenzentren; d.h., die Archivbestände des BAS würden an ein anderes CLARIN B Centre übertragen und von dort weiter nach denselben Prinzipien vorgehalten und zur Verfügung gestellt, aber unter Umständen nicht mehr aktiv gepflegt werden.

Wir hoffen, Sie als Sammlungsinhaber von der Sinnhaftigkeit einer Langzeitarchivierung am BAS überzeugt zu haben. Sollten Sie Fragen haben oder ein persönliches Gespräch suchen, setzen Sie sich bitte mit uns in Verbindung.

Florian Schiel & Christoph Draxler
URL: http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B
Email: bas@bas.uni-muenchen.de