Nutzerunterstützung und neueste Entwicklungen in Forschungsdatenrepositorien für audiovisuelle (Sprach-)Daten

Jonathan Blumtritt (jonathan.blumtritt@uni-koeln.de), Universität zu Köln, Deutschland und Felix Rau (f.rau@uni-koeln.de), Universität zu Köln, Deutschland

1 Beschreibung

Digitale Spracharchive sind ein integraler Bestandteil der Forschungsdateninfrastruktur und haben den spezifischen Auftrag audiovisuelle Sprachdaten und Dokumente zu sichern und auf deren Basis Wissensgenerierung zu ermöglichen und zu unterstützen. Ein Spracharchiv ist in diesem Sinn eine Plattform, die zwischen Produzenten und Konsumenten von Primärdaten vermittelt, so dass diese direkt oder indirekt interagieren können. Den datenproduzierenden Forschern ermöglicht das Archiv, Audio- und Videoaufnahmen menschlicher Kommunikation zu archivieren und idealerweise web-basiert zugänglich zu machen. Auf der anderen Seite werden Forscher, Sprachgemeinschaften und die weitere Öffentlichkeit in die Lage versetzt, diese Daten aufzufinden, zu betrachten, herunterzuladen und weiterzuverwenden und auf dieser Grundlage neues Wissen zu generieren. Um diesen Austausch zu unterstützen, haben die verschiedenen Spracharchive komplexe Webplattformen entwickelt.

Wie alle Forschungsdatenarchive profitieren Repositorien für audiovisuelle Daten von einem voranschreitenden Standardisierungsprozess. Das OAIS Referenzmodell hat die Grundlage für ein gemeinsames Beschreibungsvokabular gelegt. Das Data Seal of Approval/CoreTrustSeal entwickelt sich zum de-facto Standard für die Zertifizierung von Forschungsdatenrepositorien in den digitalen Geisteswissenschaften. Auf dem Gebiet der Sprachressourcen wirken die Infrastrukturinitiative CLARIN und der vom BMBF geförderte deutsche Partner CLARIN-D als starke integrative Kraft und haben mit der Etablierung von Standards, die in allen Aspekten des Datenlebenszyklus zur Anwendung kommen, gemeinsame Lösungen geschaffen. Forschungsdatenrepositorien für Sprachdaten sind herausgefordert, gültige Standards zu implementieren und gleichzeitig attraktive Dienste anzubieten, die die Spezifika der Datentypen und die Bedürfnisse der jeweiligen Nutzergruppen berücksichtigen, um eine erfolgreiche Nachnutzung von Forschungsdaten zu befördern.

Der Workshop soll Archivbetreibern, Datenkuratoren, Datenproduzenten und Datenkonsumenten die Möglichkeit zum Austausch über Angebote, Bedarfe und zentrale Weiterentwicklungen geben. Der Workshop richtet sich an Betreiber von Forschungsdatenrepositorien mit audiovisuellen (Sprach-)Daten sowie Mitarbeiter von Institutionen, die Forschungsdatenmanagement für Forschung mit audiovisuellen Daten

anbieten. Ebenso sind auch Wissenschaftler als aktive oder potentielle Nutzer von Forschungsdatenrepositorien angesprochen.

2 Ablauf

Der Workshop wird eingeleitet durch eine 15-minütige Einführung in den Themenkomplex und eine Vorstellung der Beitragenden aus verschiedenen Institutionen, die sich mit den Problemen und Lösungen rund um die Archivierung und Bereitstellung von AV-Daten auseinandersetzen.

Der Hauptteil des Workshops gliedert sich in drei Sektionen, die unterschiedliche Aspekte der Nutzerunterstützung im Datenlebenszyklus behandeln. In jeder Sektion wird es zwei aufeinanderfolgende 15-minütige Impulsvorträge geben, die diesen Themenkomplex jeweils aus der Sicht eines der beitragenden Forschungsdatenrepositorien beleuchten und Antworten darauf geben, mit welchen Maßnahmen, Policies oder technischen Implementierungen sie den jeweiligen Aspekt adressieren und wie sie diese an ihre Nutzerschaft kommunizieren. Jede Sektion endet jeweils in einer 30-minütigen Diskussion.

Der Workshop schließt mit einer kurzen Zusammenfassung und Abschlussdiskussion.

2.1 Sektion 1: Wie komme ich an die Daten?

Die Bereitstellung der Daten gehört neben einer nachhaltigen Sicherungsstrategie zu den grundlegendsten Aufgaben eines Datenarchivs. Dennoch ist die Implementierung und Vermittlung der dazugehörigen Prozesse alles andere als trivial: Erschließungsmechanismen müssen je nach Fachdomäne und Zielgruppe unterschiedlichen Anforderungen genügen, die Vergabe von Persitent Identifiern (PIDs) garantieren eine stabile Adressierung und werten die Zitierfähigkeit der Bestände auf, OAI-PMH und andere technische Schnittstellen ermöglichen eine weitergehende Auffindbarkeit in externen Portalen und Integration mit externen Diensten. Schließlich wird in dieser Frage auch der sensible Aspekt der Authentifizierung, Autorisierung und Lizenzierung berührt: Welche Bedingungen sind an den Zugriff der Daten geknüpft? Welche Verwertungsrechte werden dem Nutzer eingeräumt? Welche konkreten Schritte muss ein Nutzer unternehmen, um Zugriff zu erlangen und wie wird der Prozess kommuniziert?

2.2 Sektion 2: Wie kommen die Daten ins Archiv?

Archive haben verschiedene Workflows für die Einreichung und Integration neuer Daten. Zum Teil werden vollständig technisch geleitete Verfahren angeboten, die es dem Produzenten ermöglichen, weitgehend autark Daten über das Archiv bereitzustellen („self-archiving"). Andere Institutionen optieren bewusst für eine intensive Begleitung und Prüfung durch einen digitalen Archivar oder Kurator. Ebenso können die Herangehensweisen an die Veränderlichkeit von Datenbeständen weit auseinandergehen. Während manche Archive ausschließlich oder bevorzugt

abgeschlossene Datensammlungen integrieren, begünstigen andere in ihren Implementierung eine laufende und teils feingranulare Aktualisierung und Erweiterung von Sammlungen („living archive").

Der Wirkungsbereich eines digitalen Archivs geht idealerweise weit über den eigentlichen Einreichungsprozess hinaus. Qualitätssicherung beginnt mit einer frühen Beratung, Begleitung/Schulung und der Etablierung von Workflows für Arbeitsgruppen, die Daten in einem Repositorium archiven wollen. Datenarchive sind nicht selten auch an der Entwicklung von Software beteiligt, die in der Korpuserstellung und -kuratierung zum Einsatz kommt (z.B. Annotations-Tools, Metadaten-Editoren).

2.3 Sektion 3: Was kann ich mit den Daten machen?

Archive können „Mehrwert-Dienste" anbieten, die die Daten in einer fachspezifischen Art und Weise darstellen, kontextualisieren und vernetzen, oder sogar die Auswertung und Weiterverarbeitung unterstützen. In welcher Weise unterstützt das Archiv den wissenschaftlichen Nutzer des Archivs?

3 Vorträge

3.1 Dynamische Forschungsdatenrepositorien für die Geisteswissenschaften

Der Vortrag präsentiert den Ansatz eines dynamischen Forschungsdatenrepositoriums für die Geisteswissenschaften, der im BMBF-Zentrumsprojekt „Kölner Zentrum Analyse und Archivierung von AV-Daten" (KA³) implementiert wird. Im Mittelpunkt steht die Vereinbarkeit von Skalierbarkeit und Berücksichtigung fachspezifischer Anforderungen.

Andreas Witt (andreas.witt@uni-koeln.de), Institut für Digital Humanities, Universität zu Köln und *Michael Lönhardt* (loenhardt@uni-koeln.de), Diensteentwicklung, Regionales Rechenzentrum, Universität zu Köln

3.2 Eines für alle - Alles für einen

Der Beitrag beleuchtet die besonderen Anforderungen, die an ressourcentyp- bzw. fachspezifische Repositorien gestellt werden. Am Beispiel des HZSK-Repositoriums am Hamburger Zentrum für Sprachkorpora wird das Spannungsfeld zwischen dem Ziel einer möglichst breiten Nutzbarkeit und der Anpassung an spezifische Nutzergruppen diskutiert.

Hanna Hedeland (hanna.hedeland@uni-hamburg.de) Hamburger Zentrum für Sprachkorpora (HZSK), CLARIN-D, Universität Hamburg und *Timm Lehmberg* (timm.lehmberg@uni-hamburg.de), INEL, Institut für Finnougristik/Uralistik, Universität Hamburg

3.3 Muss es immer Fedora sein? Die Repositoriumslösung der Sprachbank von Finnland (Kielipankki)

Vorgestellt wird ein alternativer Ansatz der Datenbereitstellug, der ohne Repositoriumssoftware wie Fedora oder DSpace auskommt. Das Zusammenspiel zwischen verschiedenen Zugangsformen, PID-Verwaltung, Metadatenverwaltung, Versionierung und Zugangskontrolle wird erläutert, wobei besonders auf den Download-Service, die PID- und Metadatenverwaltung näher eingegangen wird.

Martin Matthiesen (martin.matthiesen@csc.fi), The Language Bank of Finland, CSC - IT Center for Science

3.4 Virtuelles An-die-Hand-nehmen: Qualitätssicherung für linguistische und kulturelle Datensammlungen

ELAR betreut weltweit intensiv Linguisten mit unterschiedlichsten linguistischen und digitalen Vorkenntnissen, um diese in die Lage zu versetzen, digitale Daten selbst zu kurieren, zu archivieren und langfristig für andere nutzbar zu machen.

Vera Ferreira (vf4@soas.ac.uk), *Sophie Salffner* (ss123@soas.ac.uk) und *Mandana Seyfeddinipur* (ms123@soas.ac.uk), Endangered Languages Archive, SOAS University of London

3.5 Visualisierung zeitalignierter Audio-Annotationen mit IIIF

Der Vortrag gibt Einblick in ein im Rahmen des Projekts KA³ entwickeltes Softwaresystem zur Visualisierung zeitalignierter Audio-Annotationen. Neben einer Live Demonstration werden die konzeptionellen Anpassungen und technischen Entwicklungen vorgestellt, die nötig waren, um die REST Standardisierungsbemühungen des International Image Interoperability Framework (IIIF Image API, IIIF Presentation API, IIIF Search API) für den Bereich zeitalignierter Audio-Annotationen nutzbar zu machen.

Jochen Graf (jochen.graf@uni-koeln.de), Mitarbeiter im Projekt KA³, Regionales Rechenzentrum, Abteilung Diensteentwicklung, Universität zu Köln

3.6 Erschließung audiovisueller Daten im AGD am Beispiel des FOLK-Korpus

Im AGD werden Varietäten- und Gesprächskorpora im Deutschen archiviert und bereitgestellt. Wir konzentrieren uns in diesem Beitrag darauf, wie Audios, Videos und Transkripte durch die DGD nutzbar gemacht werden.

Jan Gorisch (gorisch@ids-mannheim.de) & *Thomas Schmidt* (thomas.schmidt@ids-mannheim.de), Programmbereich Mündliche Korpora, Institut für Deutsche Sprache, Mannheim

4 Liste der Beitragenden

Dr. Vera Ferreira, *Endangered Languages Archive, SOAS, University of London*

Vera is a trained linguist with a background in language documentation and field research. Her main research interests lie in European endangered languages and in the connection between documentary data and language revitalisation. She is head of CIDLeS (Interdisciplinary Centre for Social and Language Documentation) and digital archivist at Endangered Languages Archive (SOAS University of London). As the digital archivist, Vera provides advice and training on all aspects of data management, metadata preparation and digital archiving.

Dr. Jan Gorisch, *Institut für Deutsche Sprache, Mannheim, Programmbereich Mündliche Korpora*

Jan Gorisch ist Mitarbeiter des Archivs für Gesprochenes Deutsch (AGD) und wirkt bei Datenübernahmen und deren Kuration mit. Er arbeitet an der Automatisierung des Workflows am AGD und entwickelt Tools zur Integration von Sprachtechnologie bei der Erschließung der Sprachdaten.

Jochen Graf, M.A., *Regionales Rechenzentrum, Diensteentwicklung, Universität zu Köln*

Jochen Graf hat einen langjährigen Hintergrund in Digital Humanities Projekten und ist Entwickler im BMBF-Zentrumsprojekt „Kölner Zentrum Analyse und Archivierung von AV-Daten" (KA^3) am Regionalen Rechenzentrum an der Universität zu Köln.

Hanna Hedeland, M.A., *Hamburger Zentrum für Sprachkorpora, CLARIN-D, Universität Hamburg*

Hanna Hedeland koordiniert als Geschäftsführerin des Hamburger Zentrums für Sprachkorpora (HZSK) die Arbeit der dort angesiedelten Infrastrukturprojekte sowie die Kooperationen mit externen Forschungsprojekten. Im Rahmen des Projekts CLARIN beschäftigt sie sich mit der Entwicklung von Standards und Best Practices sowie entsprechenden Technologien und Workflows für die Aufbereitung und Bereitstellung (mehrsprachiger) gesprochener Daten.

Timm Lehmberg, M.A., *Institut für Finnougristik/Uralistik, Langzeitprojekt INEL/Hamburger Zentrum für Sprachkorpora (HZSK), Universität Hamburg*

Timm Lehmberg ist technischer Koordinator des Langzeitvorhabens INEL („Grammatiken, Korpora und Sprachtechnologie für indigene nordeurasische Sprachen"), das im Rahmen des gemeinsam von Bund und Ländern finanzierten Akademieprogramms durchgeführt wird sowie Mitwirkender am Hamburger Zentrum für Sprachkorpora (HZSK).

Dipl.-Inf. Michael Lönhardt, *Regionales Rechenzentrum, Diensteentwicklung, Universität zu Köln*

Michael Lönhardt ist Leiter der Diensteentwicklung am Regionalen Rechenzentrum der Universität zu Köln.

Martin Matthiesen, M.A., *Senior Application Specialist, CSC - IT Center for Science, Finnland*

Martin Matthiesen ist Administrator der Sprachbank von Finnland (www.kielipankki.fi/languagebank/) und beschäftigt sich mit der nachhaltigen Bereitstellung von multimodalen Forschungsdaten für die Sprachwissenschaft und angrenzenden Disziplinen. Dies umfasst Versionierung, Metadaten, Persistente Identifikatoren (PIDs) und Zugangsverwaltung für nicht frei zugängliche Daten. Die Sprachbank ist ein gemeinsamer Service von CSC - IT Center for Science (www.csc.fi) und der Universität Helsinki (www.helsinki.fi/en) für FIN-CLARIN (www.kielipankki.fi/organization/).

Dr. Sophie Salffner, *Endangered Languages Archive, SOAS, University of London*

Sophie Salffner ist Digitale Archivarin am Endangered Languages Archive an der SOAS University of London. Sie arbeitet u.a. in der Aus- und Weiterbildung der BenutzerInnen des Archivs und schult WissenschaftlerInnen im wissenschaftlichen Datenmanagement, in der Aufbereitung von Metadaten und im Archivieren in digitalen Archiven.

Dr. Thomas Schmidt, *Institut für Deutsche Sprache, Mannheim, Programmbereich Mündliche Korpora*

Thomas Schmidt leitet das Archiv für Gesprochenes Deutsch (AGD) und den Aufbau des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) am IDS. Mit EXMARaLDA und der Datenbank für Gesprochenes Deutsch (DGD) beschäftigt er sich außerdem mit der Entwicklung von Technologie für die Arbeit mit und den Zugriff auf audiovisuelle Daten gesprochener Sprache.

Dr. Mandana Seyfeddinipur, *Endangered Languages Archive, SOAS University of London*

Mandana Seyfeddinipur ist Leiterin des Endangered Languages Archive (ELAR) an der SOAS University of London.

Prof. Dr. Andreas Witt, *Institut für Digital Humanities/Data Centre for the Humanities, Universität zu Köln*

Andreas Witt ist geschäftsführender Direktor des Instituts für Digital Humanities an der Universität zu Köln. Er beschäftigt sich mit digitalen Forschungsinfrastrukturen für linguistische Ressourcen, insbesondere mit der Standardisierung von Datenformaten und mit ethischen und juristischen Aspekten beim Umgang mit Forschungsdaten.

5 Ausrichter des Workshops

Der Workshop wird von dem an der Universität zu Köln angesiedelten BMBF-Zentrum „Kölner Zentrum Analyse und Archivierung von AV-Daten" (KA^3) ausgerichtet. Das Zentrum wird am Standort Köln vom Institut für Liguistik (IfL), dem Regionalen Rechenzentrum der Universität zu Köln (RRZK) und dem Data Center for the Humanties (DCH) getragen.