

Leben gelabelt

Computergestützte Inhaltverzeichnisse für Oral-History-Interviews

Philipp Bayerschmidt und Dennis Möbus

1. Einleitung

Von der fortschreitenden Digitalisierung werden zunehmend qualitative Quellenbestände erfasst, so auch Oral-History-Interviews. Im Rahmen des DFG-Forschungsprojekts Oral-History.Digital (oh.d) ist eine Erschließungs- und Rechercheplattform für Oral-History-Interviews entstanden, die Ende 2023 freigeschaltet wurde.¹ Alle Interviews sind mit Metadaten erschlossen, die meisten transkribiert. Für die Recherche können Metadaten, Registereinträge und die Transkripte über eine Volltextsuche genutzt werden. Schlagworte auf Interviewebene oder Inhaltsverzeichnisse sind dagegen nicht bei allen Sammlungen zu finden. Da sich aus dem Namen der Sammlung, der ein Interview zugeordnet ist, oder der ursprünglichen Fragestellung, zu der ein Interview geführt wurde, nicht der vollständige Inhalt der äußerst vielschichtigen und komplexen lebensgeschichtlichen Interviews ableiten lässt, würde die Bereitstellung von Inhaltsverzeichnissen die Recherche nach thematisch passenden Interviews deutlich verbessern. Doch nur einige der auf oh.d bereitgestellten Sammlungen sind in dieser Weise aufwändig manuell erschlossen worden. Der vorliegende Aufsatz dokumentiert das Training eines Topic Models – eine Methode des maschinellen Lernens – zur automatisierten Erstellung von Inhaltsverzeichnissen für alle in oh.d bereitgestellten Interviews (Hodel et al. 2022). In der Umsetzung galt es, eine ganze Reihe von Herausforderungen zu bewältigen, die im Folgenden transparent dargestellt und diskutiert werden sollen.

Zunächst (1.) werden das oh.d-Portal und das für das Training des Topic Models daraus gewonnene Korpus (Stand Mai 2024) vorgestellt. Anschließend (2.) wird die Anwendung verschiedener Verfahren zur automatischen Inhaltserschließung unstrukturierter Textdaten auf das oh.d-Korpus diskutiert. Für die letztlich vielversprechendste Methode, das Topic Modeling mit *Latent Dirichlet Allocation* (LDA) in der Java-Implementierung MALLET ist ein umfangreiches Pre-Processing erforderlich, das in einschlägigen Vorstudien entwickelt wurde (Hodel et al. 2022; Franken/Möbus 2024; Bayerschmidt et al. im Erscheinen). Die durchgeführten Schritte werden ausführlich dargestellt und evaluiert (3.). Dafür schlagen wir eine verfeinerte qualitative Evaluationsroutine für das Topic Modeling vor, die auf *Scalable Readings* basiert. Die Topic-Listen lassen wir abschließend von einer Gruppe Oral Historians labeln, wodurch zugleich die Qualität der Topics geprüft wird, um diese als Inhaltsverzeichnis nutzen zu

¹ Oral-History.Digital. Interviewportal – Erschließungsplattform – Forschungsumgebung, <https://portal.oral-history.digital/de>; Projektwebseite: <https://www.oral-history.digital> (14.9.2025).

können. Der Aufsatz schließt mit einem Einblick in die Implementierung des auf Topic Modeling basierenden Registers und der Inhaltsverzeichnisse in die oh.d-Plattform (4.).

2. Das Korpus aus Oral-History.Digital

Mit dem Release des Online-Portals Oral-History.Digital im September 2023 steht Forschenden eine zentrale Rechercheplattform für Oral-History-Interviews zur Verfügung. Über drei Jahre wurde, gefördert von der DFG und in enger Abstimmung mit Sammlungsinhaber*innen, oh.d als Erschließungs- und Recherche-Plattform für wissenschaftliche Sammlungen audiovisueller narrativer Interviews entwickelt.² Für die in oh.d eingestellten Sammlungen bleiben die jeweiligen Sammlungsinhaber*innen verantwortlich. Unterstützt werden sie dabei durch ein differenziertes Rechtmanagement. Die Video- oder Audiointerviews können mit Metadaten, Untertiteln, Transkripten und weiteren Begleitmaterialien versehen werden. Für die Recherche in den Sammlungen stehen verschiedene Werkzeuge wie Metadatenfilter- und Volltextsuche sowie samm-lungsübergreifende Suchoptionen zur Verfügung (vgl. ausführlich Pagenstecher 2024).

Für die Zusammenstellung des Korpus für das Topic Modeling wurde die oh.d-Plattform genutzt. Leitendes Interesse bei der Auswahl der Interviews war einzig, die thematische Breite der Sammlungen in oh.d abzubilden. Eine inhaltliche Vorgabe bestand nicht. Ausgewählt wurden insgesamt 984 Interviews unter der Vorstellung, dass annähernd 1.000 Interviews bzw. über 20 Millionen Wörter eine sinnvolle Trainingsgröße darstellen. Die Interviews stammen aus Beständen von sieben in oh.d vertretenen Archiven: dem Archiv „Deutsches Gedächtnis“ des Instituts für Geschichte und Biographie der FernUniversität in Hagen, den Archiven „Zwangsarbeit 1939-1945“ und „Colonia Dignidad. Ein chilenisch-deutsches Oral History-Archiv“ der Freien Universität Berlin, der Werkstatt der Erinnerung der Forschungsstelle für Zeitgeschichte in Hamburg, dem Zeitzeugenarchiv des Dokumentationszentrum Flucht, Vertreibung, Versöhnung, der Sammlung „Erzählte Lebensgeschichte“ des Museum Friedland und einer Sammlung des Hannah-Arendt Instituts zur Erziehung in der DDR. Aufgrund der Funktionsweise von Topic Modeling konnten nur deutschsprachige Interviews Verwendung finden. Für die Nutzung der Interviews als Trainingsdaten konnten wir vollständige Transkripte nutzen.

3. Diskussion der Verfahren zur automatischen Inhaltserschließung

Zur automatischen Inhaltserschließung unstrukturierter Textdaten gibt es zahlreiche Verfahren (vgl. auch vor dem Hintergrund sensibler Daten den noch aktuellen Überblick bei Schöch et al. 2020). Das Ziel von Topic Modeling, das dem maschinellen Lernen zuzuordnen ist, ist es, themenähnliche Wortlisten aus Textkorpora auf Grundlage von Wahrscheinlichkeitsrechnung zu erstellen, die im besten Falle einer semantischen Wortliste entsprechen. Verfahren, die dem Topic Modeling zuzurechnen sind, zeichnen sich dadurch aus, dass sie nicht auf punktuelle Information abzielen. Das unterscheidet sie von anderen unüberwachten Verfahren (die explorativ arbeiten, wie etwa Wortfrequenzanalysen) sowie überwachten Verfahren (wie einer Named Entity Recognition (NER), die nach vorgegebenen Kategorien und Vokabeln wie Namen und Orten sucht). Seit den 1990er Jahren wurden verschiedene Topic-Modeling-Verfahren und -

2 Siehe: <https://www.oral-history.digital/projekt/index.html> (9.7.2024)