VOM WERT DIGITALER TONDOKUMENTE FÜR DIE HISTORISCHE FORSCHUNG

Was sind „Forschungsdaten"? Handelt es sich lediglich um einen weiteren naturwissenschaftlichen Begriff, der ohne Rücksicht auf seinen Entstehungszusammenhang nun in die Geisteswissenschaften getragen wird? Oder sprechen wir von einem Gegenstand der Forschung, dessen Bedeutung in Zeiten wachsender Datenmengen – eben auch in den Geisteswissenschaften – noch nicht eindeutig erkannt wurde? Bezeichnet man damit jenes Kollateralmaterial der Forschung, das nach dem Abschluss von Dissertationen, Habilitationen und Forschungsprojekten fortan ebenfalls publizistisch verwertet werden muss? Oder sind vielmehr neue Konzepte beim Entwurf von Forschungsprojekten gefragt?

Diese rege Debatte läuft schon eine geraume Weile; zusätzlich vorangetrieben wird sie durch die immer präziseren Auflagen großer Förderinstitutionen zum Umgang mit solchen Daten. Besondere Schwierigkeiten zeigen sich dabei im Umgang mit Audiodaten. Diese sind in der historischen Forschung verbreitet, werden aber längst nicht so systematisch beachtet wie in anderen Disziplinen, insbesondere den mit empirischer Datenerhebung stärker vertrauten Sozialwissenschaften.

Um diese Problematik zu diskutieren, veranstaltete das Collegium Carolinum gemeinsam mit der Graduiertenschule für Ost- und Südosteuropastudien am 11. Februar 2016 den Workshop „Vom Wert digitaler Tondokumente für die historische

Forschung", der bereits im Vorfeld großes Interesse hervorrief. Ziel war es, herauszufinden, was in der historischen Forschung hinsichtlich digitaler Tondokumente technisch möglich, rechtlich zulässig und von Förderinstitutionen erwünscht ist. Dazu sollte auch die Perspektive verwandter Fachdisziplinen auf vergleichbare Daten berücksichtigt werden.

Wie schwer diese Fragen auf einmal zu beantworten sind, machten die Veranstalter, Arpine Maniero und Johannes Gleixner vom Collegium Carolinum, in ihrer Einführung deutlich: So sei die Rechtslage bei der Publikation audiovisueller Forschungsdaten insofern unbefriedigend, als die Forscher – abgesehen von restriktiven, in der Praxis aber insuffizienten Lösungen – keine klare Handreichung hätten. Außerdem könne man in der Praxis der historischen Forschung nach wie vor nur selten eine Wirkung der zahlreichen Angebote und Möglichkeiten zum Datenmanagement feststellen.

Die Position der Deutschen Forschungsgemeinschaft (DFG) präzisierte im ersten Vortrag Stefan Winkler-Nees (Bonn), der auf die laufende Debatte um Forschungsdaten einging. Die etablierten und funktionierenden Definitionen aus den Naturwissenschaften seien auf die Forschungspraxis der Geisteswissenschaften nicht ohne weiteres übertragbar. Aber auch die DFG könne hier keine allgemeingültige Lösung anbieten, vielmehr seien die einzelnen Disziplinen gefragt, passende Vorgangsweisen zu entwickeln. Deren Fachkollegien entschieden schließlich auch über die Förderwürdigkeit von Forschung. Ebenso wie die inhaltliche Begutachtung erfolge letztlich auch die Überprüfung des Umgangs mit Forschungsdaten nach den Regeln der eigenen Disziplin.

Winkler-Nees betonte die Notwendigkeit, auf der politischen wie der fachwissenschaftlichen Ebene ein Bewusstsein für digitale Forschungsinfrastrukturen zu schaffen. Gewinnbringend – aber auch unabdingbar – sei die enge Zusammenarbeit zwischen den jeweiligen Disziplinen und der Fachinformatik, aber auch die Kooperationen mit den vorhandenen Forschungsinfrastrukturen, wie etwa dem Serviceprojekt Informationsinfrastruktur (INF),[1] das in Freiburg entwickelt wird. Solche Kooperationen machten deutlich, dass gegenwärtig nicht mehr der Aufbau einer Informationsdatenstruktur, sondern die Weiterentwicklung des Bestehenden die wichtigste Aufgabe sei, wobei es darum gehen müsse, Parallelentwicklungen zu vermeiden. Als eines der größten Probleme bezeichnete er allerdings die institutionelle Finanzierung von digitalen Forschungsinfrastrukturen, die – anders als die der institutseigenen Bibliotheken, für die im Normalfall ein festes Budget vorgesehen sei – völlig fehle.

Die DFG könne hier nur Empfehlungen aussprechen, wie etwa im Positionspapier der AG Forschungsdaten der Allianz der Wissenschaftsorganisationen. Diese Vorgaben könnten lediglich als Rahmen dienen. Für die Förderrichtlinien unterstrich Winkler-Nees, dass die Förderpraxis klar zwischen Datenmanagement (Infrastrukturförderung) und Datennutzung (Sachförderung) unterscheide, wobei IT-Projekte eher in die erste und wissenschaftliche in die zweite Kategorie fallen. In

[1] Serviceprojekt Informationsinfrastruktur. URL: https://www.sfb1015.uni-freiburg.de/info (letzter Zugriff 19.04.2016).

jedem Fall sei es schon heute sinnvoll, einen Datenmanagementplan bei Projektanträgen zu berücksichtigen bzw. entsprechende Kooperationen mit Informationsstrukturbetreibern rechtzeitig zu bedenken.

Daraus lässt sich der vorläufige Schluss ziehen, dass der mit Forschungsdaten planende Forscher im Moment sowohl Rahmenförderbedingungen, die einen Datenmanagementplan nahelegen, als auch die möglicherweise abwehrende Reaktion der eigenen disziplinären Zunft berücksichtigen muss. Im Gegensatz zu Linguisten und Sozialwissenschaftlern befinden sich die historischen Wissenschaften also erst am Beginn einer Übergangsphase. In absehbarer Zeit könnte die Vorlage von Datenmanagementplänen zum unumgänglichen Standard werden.

Im Anschluss an Winkler-Nees sprach Thomas Schmidt (Mannheim), der am Institut für deutsche Sprache (IDS) den Programmbereich „Mündliche Korpora" leitet, über den Stand der Entwicklung am „Archiv für gesprochenes Deutsch" (AGD).[2] Die dort entwickelte Datenbank für gesprochenes Deutsch (DGD) stellt eine Korpusplattform dar, deren Audiodaten mit einem Text-Ton-Alignement versehen werden, sodass die Anzeige die abgespielten Audiodaten „mitliest". Über die DGD werden diese Daten der wissenschaftlichen Öffentlichkeit verfügbar gemacht. Mit der umfangreichen Datenbank des Instituts arbeiten, so Schmidt, mittlerweile etwa 5000 registrierte Nutzer – Studierende wie Wissenschaftler – aus unterschiedlichen Fachrichtungen. Obwohl sich die im AGD befindlichen Materialien vor allem für die linguistische Forschung eignen, wurde anhand der Beispiele deutlich, dass die linguistischen Untersuchungen auch für die historische Forschung von großem Wert sein können. Die Grundvoraussetzung dafür, dass solche Bestände nachnutzbar sind, ist allerdings die konsequente Einhaltung verbreiteter Erschließungsstandards.

Aus seiner langjährigen Praxis der Einwerbung und Übernahme archivierter Sprachkorpora gab Schmidt einige praktische Hinweise für Audioaufnahmen in Forschungsprojekten. Die heutige Technologie erleichtere die Vorbereitung zur Archivierung und strukturierten Aufbereitung der Daten, da man nunmehr auch unkomprimierte Datenformate mit hohen Abtastraten (Samplingraten) schon bei der Aufnahme erzeugen und speichern könne. Auch mit Blick auf die spätere Diskussion warb Schmidt bei der Datenbearbeitung für Programme, die offenen Standards genügen, so etwa die vom IDS mitentwickelten „EXMARaLDA" und „FOLKER". Solche Überlegungen müssten Teil eines Projektdatenplans sein, da sie die Langzeitarchivierung sicherstellten. Nichtstandardisierte Sprachkorpora seien nur mit erheblichem Aufwand in die Plattformen zu integrieren.

Astrid Schoger (München) von der Bayerischen Staatsbibliothek (BSB) gab im Anschluss einen Einblick in die Praxis und Zukunft der Langzeitarchivierung von Dokumenten aller Art. Dabei sprach sie allgemeine Probleme an, die sich bei der Langzeitarchivierung stellen, beispielsweise die begrenzte Haltbarkeit der Datenträger, technologischer Wandel, kurzlebige Produktions-, Verwaltungs- und Abspielumgebungen, veraltende Dateiformate, rasant wachsende Datenmengen, zunehmende Komplexität der Daten und deren Vernetzung, aber auch das fehlende

[2] Archiv für gesprochenes Deutsch. URL: http://agd.ids-mannheim.de/index.shtml(letzter Zugriff 19.04.2016).

Problembewusstsein. Die Langzeitarchivierung soll diesen Schwierigkeiten auf verschiedenen Wegen entgegenwirken: über die Digitalisierung in der höchstmöglichen Qualität, die Nutzung standardisierter Identifikatoren wie URNs, aber auch über Qualitätskontrollen. Die BSB bietet außerdem unter der Rubrik „Daten für die Forschung" Wissenschaftlern den Zugriff auf hochauflösende Dokumente an. Momentan erstreckt sich dieses Angebot allerdings noch nicht auf Audiodaten.

Florian Schiel (München), mitverantwortlich für das Bayerische Archiv für Sprachsignale (BAS) am Institut für Phonetik und Sprachverarbeitung der LMU München, präsentierte im ersten Vortrag des Nachmittags dessen technische Möglichkeiten und Dienstleistungen. Wie das IDS zählt auch das BAS zu den Servicecentern der CLARIN-D-Infrastruktur und konzentriert sich unter anderem auf die Alignierung von Tonspuren mit Text. Schiel führte an einem konkreten Beispiel vor, wie die Hilfsprogramme WebMAUS (vollautomatische Segmentierung und Auszeichnung von Audiodateien anhand einer orthografischen Transkription), WebMINNI (Automatische phonetische Segmentierung und Auszeichnung für mehrere Sprachen ohne Text-Input) sowie ChunkPreparation (Erstellung von BAS-Partitur-Dateien mit einem Speech Chunk Tier (TRN) aus verschiedenen Eingabeformaten) funktionieren.[3] Die Nutzung dieser Programme ist plattformunabhängig und daher für alle Interessierten möglich. Trotz bestehender technischer Grenzen bestätigte dieses Beispiel ebenso wie die Ausführungen von Thomas Schmidt zuvor eine zentrale Aussage von Stefan Winkler-Nees: Die Förderung und Bereitstellung einer digitalen Infrastruktur für Forschungsdaten ist bereits weit fortgeschritten. Die Aufgabe besteht nun darin, eine verbreitete Nutzung der Dienste zu erzielen. Der Infrastrukturdienst und das wissenschaftliche Nutzerverhalten scheinen aber insgesamt eine kritische Schwelle überschritten zu haben. Schiel selbst gab ebenfalls Einblicke in das Nutzungsverhalten der Wissenschaftler, die eine Vorliebe für bestimmte Tools entwickelten, andere wiederum weitgehend ignorierten. Deswegen appellierten die Entwickler einstimmig an die Fachwelt, Dienste, die etwa im Rahmen von CLARIN-D zur Verfügung stehen, zu nutzen und vor allem Rückmeldungen zu geben, damit sich die weitere Entwicklung besser steuern lässt.

Zum Abschluss folgten zwei Berichte aus der Praxis der Sprachforschung. Zunächst sprach Klaas-Hinrich Ehlers (Berlin/München) über seine Erfahrungen bei seiner kontaktlinguistischen Untersuchung unter Heimatvertriebenen in Mecklenburg, die am Collegium Carolinum angesiedelt ist und von der DFG gefördert wird. Für dieses Projekt wurde umfangreiches Audiodatenmaterial erstellt, dessen Erhebung aber, so Ehlers, nicht mit den zuvor empfohlenen Programmen und in den gewünschten Formaten erfolgen konnte. Dazu kommen Schwierigkeiten wie etwa die Kompatibilität der Betriebssysteme auf den Rechnern angestellter Hilfskräfte bzw. menschliche Fehlinterpretationen bei der Erhebung von Audiodaten vor allem bei kleineren Projekten. Diese Probleme sind auch durch neue Technik nicht zu beheben. Erschwerend hinzu kommt die Frage des Datenschutzes: Um dem Persönlichkeitsrecht Genüge zu tun, müssten Interviewdaten – je nach Kreis der Befragten

[3] URL: https://clarin.phonetik.uni-muenchen.de/BASWebServices/#/services (letzter Zugriff 19.04.2016).

– in einem so hohen Maße anonymisiert werden, dass nur noch verwertbare Metadaten übrig blieben. Ehlers betonte, dass auch ein Datenmanagementplan angesichts dieser in der Forschung üblichen Bedingungen nicht so umgesetzt worden wäre, wie ursprünglich geplant. Damit kehrte die Diskussion zu der zu Beginn des Workshops gestellten Frage zurück, ob überhaupt mehr als nur Förderrahmenbedingungen formuliert werden können.

Im Anschluss stellte Stephan Lücke (München) von der IT-Gruppe Geisteswissenschaften der LMU das Projekt „Audioatlas siebenbürgisch-sächsischer Dialekte" vor, das eine umfangreiche Tondokumentation zugänglich gemacht hat.[4] Dieser Audioatlas liegt vollständig als Weboberfläche vor und beruht auf einem älteren Sprachkorpus, das im Zuge des Projekts transkribiert und digital aufbereitet wurde. Das Material ist nach den Kriterien Ort, Jahr, Alter, Subcorpus und Inhalt im Gesamtbestand recherchierbar und mit einzelnen Dateien verlinkt. Für die inhaltliche Analyse wurden zudem unterschiedliche Methoden angewandt. Neben diesem werden auch andere Projekte von der ITG selbstständig langzeitgesichert, so dass weitere Datenaufnahmen auch nach Projektende möglich und erwünscht sind. An der Frage, welches Datenformat für eine solche langfristige Speicherung ideal ist, entzündete sich dann eine kontroverse Debatte.

Um die Nachhaltigkeit ging es auch in der Abschlussdiskussion. Die Entscheidung, ob die Datensicherung in relationalen Datenbanken bzw. mit einer XML-Strukturierung erfolgt, oder mit welchen Datenträgern und Formaten (wav oder mp3) die Forscher arbeiten, kann für die Nachhaltigkeit erzeugter Dokumente gravierende Folgen haben. Ein generelles Problem, für das zumindest mittelfristig Lösungsansätze gefunden werden müssen, stellt nach allgemeiner Auffassung die rechtliche Situation dar. Die Forscher vernachlässigen diese Fragen oft angesichts einer komplizierten und unklaren Rechtslage. Eine nachträgliche Rechteklärung wiederum erweist sich in den meisten Fällen als schwierig bzw. nicht mehr möglich. Darunter leidet vor allem die Veröffentlichung bzw. Zugänglichkeit der Forschungsdaten.

Als vorläufiges Fazit lässt sich festhalten: Die Frage nach dem Umgang mit den Daten, die im Zuge von Forschung generiert werden, erfordert in naher Zukunft auch von den Historikern eine Antwort. Diese Antwort muss nicht eindeutig ausfallen, sollte aber die Vorzüge von Standardisierungen, insbesondere bei Audiodaten, weitestgehend berücksichtigen.

[4] Audioatlas siebenbürgisch-sächsischer Dialekte. URL: http://www.asd.gwi.uni-muenchen.de/ (letzter Zugriff 19.04.2016).

München Johannes Gleixner / Arpine Maniero