Fachwissenschaftliche Nutzungsszenarien der CLARIN-D Infrastruktur

Wiedemann, Gregor
gregor.wiedemann@uni-leipzig.de
Universität Leipzig

Gloning, Thomas
thomas.gloning@germanistik.uni-giessen.de
Universität Gießen

Blätte, Andreas
andreas.blaette@uni-due.de
Universität Duisburg/Essen

Keller, Maret
keller@gei.de
Georg-Eckert-Institut Braunschweig

Haaf, Susanne
haaf@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Würzner, Kay-Michael
wuerzner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

CLARIN-D: Eine Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften

Auffinden, Auswerten, Aufbewahren: Die Forschung mit digitalen Sprach- und Textdaten stellt die Sozial- und Geisteswissenschaften vor neue, fächerübergreifende Herausforderungen:

- Rohdaten und Ergebnisse sollen für Nachvollziehbarkeit und weitere Analysen zentral und einfach auffindbar sein,
- für die Auswertung können computergestützte Werkzeuge vielfältig eingesetzt werden – möglichst entlang methodischer Standards, und
- Forschungsergebnisse müssen nachvollziehbar und langfristig verfügbar gemacht werden.

Die Forschungsinfrastruktur CLARIN-D (Common Language Resources and Technology Infrastructure in Deutschland) unterstützt die Geistes- und Sozialwissenschaften dabei, ihre digitalen Ressourcen in nachhaltiger, offener und interoperabler Weise zur Verfügung zu stellen (www.clarin-d.de). Über neun CLARIN-D-Zentren mit unterschiedlichen Arbeitsschwerpunkten und einem breiten Angebot an Webservices stehen der deutschen Forschungslandschaft umfangreiche Angebote zur Forschung mit Sprachdaten zur Verfügung. Insofern bei den Angeboten die Perspektive von nicht computerlinguistisch vorgebildeten Fachwissenschaftler_innen als Zielgruppe im Vordergrund steht, sind die CLARIN-D Angebote für die Digital Humanities von besonderem Interesse, zielen sie doch darauf ab, die Arbeit mit digitalen Sprachdaten durch Vereinheitlichung von Standards und Bereitstellung von Werkzeugen und Webservices zu erleichtern.

Beim Auffinden von Ressourcen geht es darum, Zugang zu Daten zu erhalten, die der wissenschaftlichen Gemeinschaft zur Verfügung gestellt wurden. Diese Daten werden zitiert und können so gefunden und in anderen Forschungskontexten und zur Reproduktion von Ergebnissen verwendet werden.

Werkzeuge zur Analyse von Forschungsdaten sind über das Web zugänglich und können dadurch ohne zusätzlichen Aufwand verwendet werden. CLARIN-D macht Werkzeuge für die Geistes- und Sozialwissenschaften verfügbar, so dass unterschiedliche Werkzeuge für neue Forschungsfragestellungen zusammen verwendet werden können.

Daten, die in Forschungsprojekten entstehen oder die anderen Forschern zur Verfügung gestellt werden, können mit Hilfe der CLARIN-D-Zentren langfristig aufbewahrt und somit archiviert werden. Sie erhalten dabei eine eindeutige Referenz und können ähnlich wie Bücher und Artikel zitiert werden. Außerdem werden dadurch Anforderungen von Förderungsorganisationen zur Vorhaltung von Daten und Ergebnissen über Projektlaufzeiten hinaus gewährleistet.

Mit diesen Services ist CLARIN-D für eine Vielzahl von Fachdisziplinen, welche sich im Bereich der Digital Humanities bewegen, von großen Interesse. Um den Anforderungen und Bedürfnissen dieser Fachcommunities gerecht zu werden, haben sich (potenzielle) Nutzer_innen der Infrastruktur innerhalb von CLARIN-D in sogenannten Fach-Arbeitsgruppen (F-AGs) organisiert. In zehn F-AGs, welche von Germanistik und anderen Philologien über diverse Teilbereiche der Linguistik bis hin zu Sozial- und Geschichtswissenschaften reichen, sind mittlerweile ca. 200 Wissenschaftler_innen organisiert, welche sich über Möglichkeiten, Bedarfe und methodische Standards bei der Arbeit mit digitalen Sprachdaten austauschen. Zudem werden im Rahmen von sogenannten “Kurationsprojekten” wichtige fachwissenschaftliche Ressourcen aufbereitet und für die Forschungscommunity zugänglich gemacht.

Im geplanten Panel sollen exemplarisch zwei Use Cases dargestellt und dokumentiert werden (Abschnitt 2). Ein weiterer Beitrag befasst sich mit den Prinzipien und Möglichkeiten der Dokumentation von fachwissenschaftlichen Nutzungsszenarien von Clarin-D-Ressourcen (Abschnitt 3).

Fachwissenschaftliche Use Cases der Infrastruktur

Das Panel “Fachwissenschaftliche Nutzungsszenarien der CLARIN-D Infrastruktur” stellt zwei Use Cases aus unterschiedlichen fachwissenschaftlichen Perspektiven vor, bei denen Services bezüglich der drei zentralen Leistungen Auffinden, Auswerten und Aufbewahren zur Anwendung kommen. Über den Erkenntnisgewinn der Darstellung der einzelnen Use Cases hinaus wird so im Rahmen des Panels der Nutzen der Infrastruktur für unterschiedliche Disziplinen insgesamt sichtbar gemacht. Dazu stellen die Einzelvorträge die Generalisierbarkeit ihrer Ansätze an zentralen Punkten heraus um zu verdeutlichen, wie andere Forschungsfragen mit ähnlichen Methoden und Werkzeugen bearbeitet werden können. Zudem wird im Rahmen der Diskussionen zu den einzelnen Vorträgen vor allem auf Erfahrungen und Generalisierbarkeit der Ansätze fokussiert, um so den projekt- und fächerübergreifenden Mehrwert der Nutzung von Komponenten der Infrastruktur aufzuzeigen und die Anwendung für eigene Projekte zu ermutigen. Vorgestellt werden drei Nutzungsszenarien aus dem Bereich der Politikwissenschaften, der Neueren Geschichte und der Germanistik, wobei neben den Projekten vor allem die Reflexion des Einflusses der CLARIN-D Infrastruktur auf Forschungsmöglichkeiten und Dokumentation von Forschungsabläufen im Vordergrund steht.

Beitrag 1: Aufbereitung und Analyse von Parlamentsprotokollen als öffentliche Sprachressource der Demokratie

Im deutschsprachigen Raum besteht ein Mangel an frei verfügbaren, annotierten Korpora politischer Texte. Dadurch wird der Einstieg in die DH-Forschung massiv gehemmt. Eine mögliche Lösung bieten Plenarprotokolle als öffentliches Archiv des politischen Zeitgeschehens einer Demokratie, welche im Rahmen der CLARIN-D Infrastruktur aufbereitet und verfügbar gemacht werden. Plenarprotokolle des Bundestags, der Landtage oder auch des Europaparlaments dokumentieren über große Zeiträume das gesamte Spektrum politischer Aktivität. Dies ist zugleich, ohne eine thematische Klassifikation von Debatten, ein Nachteil: Plenarprotokolle decken, wenn nicht Subkorpora nach inhaltlichen Kriterien gebildet werden können, das politische Geschehen für viele Auswertungszwecke zu undifferenziert ab. Für eine Vielzahl sozialwissenschaftlicher Fragestellungen ist es relevant, dass themen- bzw. politikfeldspezifische Subkorpora gebildet werden können. Dafür ist eine Annotation von Debatten und deren Klassifikation erforderlich. Eine entsprechende Aufwertung der Ressource “Plenarprotokollkorpus” wird im Kurationsprojekt der F-AG 8 vorgenommen.

Im Rahmen des Panel-Beitrags wird zuerst der Workflow dargestellt, wie auf Basis der bei Parlamenten verfügbaren Plenarprotokolle (im txt- und pdf-Format) XML-Dokumente aufbereitet werden, die TEI-Standards entsprechen. Anschließend zeigen wir, wie mit CLARIN-Tools Annotationen für die Korpusaufbereitung vorgenommen werden. Redebeiträge sollen auf Basis der manuellen Annotation automatisch in bestimmte Themenkategorien klassifiziert werden. Im Beitrag wird auch dargestellt, wie die Qualitätssicherung der (halb-)automatischen Aufbereitung durch Intercoderreliabilitäten und Qualitätskriterien maschineller Sprachverarbeitung von Seiten der klassischen Politikwissenschaft als auch von Seiten der Informatik realisiert werden kann. Versionierung der Daten und Issue Tracking werden dabei gleichermaßen realisiert. Im Ergebnis steht der sozialwissenschaftlichen Community (und selbstverständlich auch anderen Wissenschaftler_innen) ein nach Themengebieten differenziert auswertbares Korpus zur Verfügung.

Beitrag 2: CLARIN-kompatible Aufwertung OCR-erfasster Texte aus der Lehrbuchsammlung des GEIs und deren Integration in die CLARIN-D-Infrastruktur: Ein Fazit aus fachwissenschaftlicher Sicht

Welchen Aufwand bringt eine Integration historischer Quellen in die CLARIN-D Infrastruktur mit sich? Welche Mehrwerte ergeben sich daraus für die historische Forschung? Diese Fragen sollte das Projekt „ Quellen des Neuen: Realkundliches und naturwissenschaftliches Wissen für Dilettanten und Experten zwischen Aufklärung und Moderne“ der 2014 gegründeten CLARIN-D-Facharbeitsgruppe „Neuere Geschichte“ klären. Zu diesem Zweck sollten im Projekt Korpora aus verschiedenen Projektkontexten über die CLARIN-Infrastruktur miteinander verknüpft und interoperabel gemacht werden. Im Blick waren dabei (1) das digitale Schulbuchkorpus des GEI (Georg-Eckert-Institut Braunschweig; GEI), (2) das Korpus des Deutschen Textarchivs (Berlin-Brandenburgischen Akademie der Wissenschaften; BBAW) sowie (3) die gedruckten Publikationen des Universitätsgelehrten Johann Friedrich Blumenbach (1752–1840) (Akademie der Wissenschaften Mainz). Durch Verknüpfung dieser Korpora sollten Untersuchungen zum Zusammenhang und Verhältnis von schulischer Lehre mit der Wissensproduktion und -vermittlung im universitären Umfeld ermöglicht werden. Die CLARIN-Infrastruktur ermöglicht zudem den Rückgriff auf gegebene Tools für die vergleichende Analyse der gegebenen Korpora.

Die Herausforderung des Projekts bestand darin, die in Qualität und Form heterogenen Daten der verschiedenen Korpora CLARIN-konform aufzubereiten und miteinander interoperabel zu machen. Von mehreren möglichen

Integrationsszenarien wurde in Zusammenarbeit mit dem CLARIN-D-Servicezentrum an der BBAW eine sehr genaue und vergleichsweise aufwändige Methode realisiert, mit dem Ziel, adäquate Forschungsdaten für die Wissensgeschichte und andere historisch arbeitende Disziplinen zu erhalten. Unter Nachnutzung bzw. Anpassung bestehender Workflows der BBAW wurden exemplarisch Schulbücher verschiedener Fächer und Zeiträume dem Auszeichnungslevel der DTA-Korpora und der bereits integrierten „Blumenbach-Online“-Ressourcen angepasst. In verschiedenen ‘Stadien’ der Textaufbereitung wurde zudem die Anwendbarkeit von CLARIN-D-Tools auf Teilmengen der verfügbaren Daten getestet.

Der Panel-Beitrag erläutert den Workflow zur CLARIN-konformen Aufbereitung der GEI-digital-Ressourcen und führt den Mehrwert dieser Arbeiten für die Forschung beispielhaft vor.

Beitrag 3: Möglichkeiten und Prinzipien der Dokumentation von fachlichen Nutzungsszenarien von Clarin-D-Ressourcen

Für die fachliche Nutzung der Ressourcen (Daten, Werkzeuge), die in einer Infrastruktur angeboten werden, ist die Frage von zentraler Bedeutung, wie man typische wissenschaftliche Anwendungsszenarien, bei denen Daten und Werkzeuge für eine spezifische fachliche Aufgabenstellung genutzt werden, so darstellen kann, dass die Dokumentationen für unterschiedliche Zielgruppen hilfreich und im besten Fall auch stimulierend sind. Dabei sind einerseits unterschiedliche Nutzertypen (z. B. Doktorand_innen, Studierende, erfahrene und weniger erfahrene Forscher_innen), andererseits die ganz unterschiedlichen fachlichen Fragestellungen in verschiedenen Disziplinen (z. B. Germanistik) und Unterdisziplinen (z. B. Syntax, Wortschatzforschung) zu berücksichtigen. Gegenstand des Panel-Beitrags sind zunächst drei Typen der Dokumentation, die jeweils von einer fachlichen Fragestellung ausgeht, sodann die Ermittlung von Daten und Werkzeugen umfasst, die zielorientierte Anwendung von Daten und Werkzeugen beschreibt und mit dem Hinweis auf ein fachliches Resultat endet, das auf die ursprüngliche fachliche Fragestellung rückzubeziehen ist. Die drei Typen sind:

- gedruckte und bebilderte Anleitungen;
- Screencasts, bei denen eine fachliche Anwendung digitaler Ressourcen von einer Stimme aus dem Off kommentiert wird;
- Experten-Interviews, die vor Ort oder als virtuelles Video-Meeting aufgezeichnet werden und dann als Filme / Podcasts dauerhaft zugänglich sind. Hier vertritt ein/e Interviewer/in die Nutzerinteressen, eine Expertin oder ein Experte stellt die Ressourcen und ihre Anwendung dar.

Der Beitrag wird zunächst die Prototypen vorstellen und dann allgemeine Überlegungen zu Zielen, Verfahrensweisen, Prinzipien (z. B. Usability, Zielgruppenorientierung), Darstellungsformen und zum Repertoire von Darstellungsmitteln anschließen.

Fächerübergreifende Perspektiven

Wir versprechen uns durch die multiperspektivische Reflexion der hier vorgestellten Nutzungsszenarien nicht nur eine Verbesserung des Verständnisses für den Nutzen einer Forschungsinfrastruktur, sondern hegen insbesondere die Hoffnung, einen wichtigen Beitrag zur Methodendiskussion in den Digital Humanities zu liefern. In der gemeinsamen Nutzung und Weiterentwicklung von Technologien zum Auffinden, Auswerten und Aufbewahren digitaler Sprachdaten liegt der Schlüssel zur Etablierung von Standards, zum Herabsetzen von Zugangsschwellen und damit letztlich zur Ermöglichung eines breiten Austausches in diesem noch vergleichsweise jungen Forschungsfeld.

Ablauf

Das Panel wird moderiert von Gregor Wiedemann, Koordinator der Fach-AGs im CLARIN-D Projekt. Der inhaltliche Ablauf wird wie folgt gestaltet:

- Einführung zur CLARIN-D Infrastruktur : Gregor Wiedemann (Universität Leipzig, Koordinator für die CLARIN-D Fach-AGs), Vortragszeit: 5 Minuten
- Beitrag 1, Vortragender: Prof. Dr. Andreas Blätte (Universität Duisburg/Essen, Lehrstuhl für Public Policy und Landeskunde), Vortragszeit: 10 Minuten
- Beitrag 2, Vortragende: Dr. Maret Keller (Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung), Susanne Haaf (Berlin-Brandenburgische Akademie der Wissenschaften), Kay-Michael Würzner (Berlin-Brandenburgische Akademie der Wissenschaften), Vortragszeit: 10 Minuten
- Beitrag 3, Vortragender: Prof. Dr. Thomas Gloning (Universität Gießen, Professur für Germanistische Sprachwissenschaft), Vortragszeit: 10 Minuten
- Panel-Diskussion zu fächerübergreifenden Problemen bei der Arbeit mit digitalen Sprachdaten und Lösungsansätzen im Rahmen der CLARIN-D Infrastruktur, Diskussionszeit: 25 Minuten
- Publikumsdiskussion Möglichkeiten, Perspektiven aber auch (Einstiegs-)Hürden bei der Anwendung virtueller Forschungsinfrastrukturen. Neben Verständnisfragen sollen vor allem Bedarfe seitens der (potenziellen) Anwender_innen und Möglichkeiten zur Generalisierung der präsentierte Forschungsabläufe diskutiert werden. Diskussionszeit: 30 Minuten