Einleitung

Über alle Fächer, Disziplinen und akademischen Kulturen hinweg gilt: Jede Art der wissenschaftlichen Arbeit an/mit einem Thema produziert Forschungsdaten.

Diese können beispielsweise die Form handschriftlicher Notizen, eines Labortagebuchs, aber auch die eines Sammelsuriums verschiedener Word-Dateien annehmen. Im Endeffekt bestehen Aufsätze, Monographien und andere akademische Publikationen auf ausgewählten und besonders aufbereiteten Forschungsdaten. In den konkreten Formen und Definitionen der ubiquitären »Forschungsdaten« spiegelt sich die Vielfältigkeit der wissenschaftlichen Disziplinen mit all ihren Gegenständen, Forschungsverfahren und Perspektiven wider. Dieser Tatsache trägt seit 2015 auch die Deutsche Forschungsgemeinschaft im Kontext ihrer Förderregularien in den 'Leitlinien zum Umgang mit Forschungsdaten' Rechnung:

> „Forschungsdaten sind eine wesentliche Grundlage für das wissenschaftliche Arbeiten. Die Vielfalt solcher Daten entspricht der Vielfalt unterschiedlicher wissenschaftlicher Disziplinen, Erkenntnisinteressen und Forschungsverfahren."
> (Senat der Deutschen Forschungsgemeinschaft, 2015: S. 1)

In diesem Sinne müssen Forscher*innen bereits während der Projektplanung und Antragsstellung zeigen, dass sie sich Gedanken über die entstehenden Forschungsdaten sowie ihre langfristige Sicherung und Bereitstellung gemacht haben. Dies ist besonders wichtig, da nicht nur die akademischen Publikationen an sich, sondern auch die Forschungsdaten des Prozesses hin zur Publikation „wichtige Anschlussmöglichkeiten für die weitere Forschung" liefern und zur Qualitätssicherung beitragen (Senat der Deutschen Forschungsgemeinschaft, 2015: S. 1).

Ein Grundsatz zeitgemäßen Forschens muss es also sein, die eigenen Forschungsdaten zu reflektieren und diese in geeigneter Form der (akademischen) Öffentlichkeit bereitzustellen. Wie die 'geeignete Form' auszusehen hat, war und ist Gegenstand von Diskussionen; als wichtige Orientierungspunkte haben sich allerdings seit 2016 die sog. FAIR-Prinzipien etabliert: *Findability*, *Accessibility*, *Interoperability* und *Reusability* (vgl. Harrower et al., 2020: S. 3).

Die ersten beiden Prinzipien — Auffindbarkeit und Zugänglichkeit — lassen sich auf technischer Ebene verhältnismäßig leicht umsetzen: Benötigt wird dafür in erster Linie »nur« ein Webserver, der das Hoch- und Herunterladen von Dateien ermöglicht. Diese Aufgabe wird meist von Repositorien übernommen, die von akademischen Institutionen oder privatwirtschaftlichen Unternehmen betrieben werden. Schwierigkeiten ergeben sich dort vor allem, wenn es um das Urheberrecht an Forschungsdaten sowie deren Lizenzierung und Verwertung geht; dieser Aspekt muss primär auf technischen, sozialen und ökonomischen sowie juristischen Ebenen verhandelt werden.
Wesentlich anders verhält es sich bei den Prinzipien Interoperabilität und Nachnutzbarkeit: Hier liegen die Herausforderungen auf metaphysisch-ontologischer Ebene, wobei »Ontologie« aus der Schnittmenge von Informatik und Philosophie als „a possible conceptualization of a part of the world" (Flanders & Jannidis, 2019: S. 181) verstanden wird. Herausforderungen

dieser Art folgen aus der Tatsache, dass jede Forschungsfrage und -methode eine eigene Modellierung des jeweiligen Gegenstands erzeugt (vgl. Thalheim & Nissen, 2015: S.615-617). So wird etwa das Modell — welches ein und dasselbe Manuskript beschreibt — aus sprach-, literatur- und geschichtswissenschaftlicher Perspektive unterschiedliche Schwerpunkte haben:

Je nach Fragestellung kann der Ausgangspunkt für eine Untersuchung aus sprachwissenschaftlicher Perspektive die ausführliche Annotation der morphologischen, syntaktischen und weiteren Eigenschaften einzelner Texteinheiten bilden, während analog dazu eine literaturwissenschaftliche Perspektive eher vom Paratext sowie von der internen und externen Intertextualität des Textes und seiner Teile ausgeht. Für eine geschichtswissenschaftliche Perspektive steht wiederum die Verknüpfung der im Text und in den Metadaten erwähnten Personen, Orte und Ereignisse mit ihren historischen Entitäten im Vordergrund. Aus diesen rein exemplarischen Unterschieden der Schwerpunktsetzung im Modellieren ein und desselben konkreten Gegenstandes ergeben sich verschiedene semantische Modelle, die einen starken Einfluss auf die technische und digitale Repräsentation des Gegenstandes haben; und damit auch auf die Beschaffenheit der Forschungsdaten für die interdisziplinäre Arbeit.

Gerade für die textorientierten Wissenschaften ergeben sich trotz aller Unterschiede große Schnittmengen: Die auffindbaren und zugänglichen sprachwissenschaftlichen Forschungsdaten über ein Manuskript stellen einen »Schatz« für literatur- und geschichtswissenschaftliche Fragestellungen dar und umgekehrt; schließlich wurde jeweils bereits »Grundlagenforschung« geleistet, auf die im Weiteren aufgebaut werden kann. Doch lässt sich nicht davon ausgehen, dass die Forschungsdaten auch mit interdisziplinären Fragestellungen kompatibel sind: Interoperabilität und Nachnutzbarkeit — vor allem in Bezug auf Speicherung und Abfrage von Daten sowie auf ihre Qualität — müssen jeweils erst hergestellt werden; sie sind nicht ohne weiteres gegeben. Diese Herausforderungen sind hochaktuell, gerade durch die Zunahme von Projekten, die große Mengen an Forschungsdaten produzieren, und die verstärkte Nutzung von Semantic-Web-Technologien. Speziell für (digitale) historische Untersuchungen steht laut Francesco Beretta die Frage im Raum: „[H]ow is it possible to adopt a quasi-universal *conceptualization*, an ontology, in order to ensure the interoperability of the data produced by historians?“ (Beretta, 2021: S.280).

Diese Frage lässt sich noch weiter fassen; sie betrifft nicht nur genuin digitale Projekte, sondern jegliche wissenschaftliche Arbeit, und das über alle Disziplinen hinweg. Zwar arbeitet nicht jede Disziplin explizit mit Daten im informationstechnologischen Sinne, aber generell gilt: „[I]n allen Wissenschaftsdisziplinen [entstehen] digitale Forschungsdaten“ (Kindling & Schirmbacher, 2013: S. 130). Beispielhaft für die Geschichtswissenschaften hat dies jüngst Boris Queckbörner untersucht (vgl. Queckbörner, 2019). Das Nachnutzen und Anreichern von Forschungsdaten besteht demnach zum überwiegenden Teil aus dem Speichern und Verarbeiten digitaler Daten; und genau das geschieht in verschiedenen Datenbanksystemen. Diese Systeme bilden das Fundament, auf dem Forschungsdaten repräsentiert, angereichert und ausgewertet werden. Gleichzeitig schränken sie aber bereits mögliche Anwendungsszenarien für die eigene Nachnutzung fremder und die fremde Nachnutzung eigener Daten ein:

Nicht jedes Fundament ist für jedes Szenario gleich gut geeignet. Die technische und semantische Umsetzung — in Datenbanksystemen — ist also eine der wichtigsten Entscheidungen für Interoperabilität und Nachnutzbarkeit.

Im Hinblick auf mögliche Datenbanksysteme leistet diese Arbeit einen Beitrag zur Frage, wie sich eine interdisziplinäre Nachnutzung und Anreicherung korpuslinguistischer Forschungsdaten konkret umsetzen lässt. Dabei stammen die Forschungsdaten aus dem Korpus *RIDGES Herbology 9.0* — Register in Diachronic German Science — dessen Fokus auf Kräuterkunde sich aus dem Korpus-Design ergibt: Zur Gewährleistung von Vergleichbarkeit wurden Texte einer wissenschaftlichen Disziplin gewählt, die im gesamten Untersuchungszeitraum auf ähnliche Weise vertreten ist. Die Version 9.0 umfasst 46 Werke, bestehend aus insgesamt 73 Texten, die sich auf die Zeit von 1482 bis 1914 verteilen (vgl. Odebrecht et al., 2020). Konkret geht es um die Metadaten der 46 Werke sowie um sämtliche annotierte Personennamen, auf die in den Werken referiert wird. Über die Verknüpfung der Orts- und Personennamen mit den dahinterstehenden Entitäten - soweit es möglich war, die Zuordnung zu disambiguieren - findet neben der Nachnutzung noch eine interdisziplinäre Anreicherung statt. Der technische Teil der Nachnutzung und Anreicherung wird mit Python-Skripten und unter Nutzung der auf MediaWiki basierenden Software-Suite *Wikibase* umgesetzt. Folgende Fragen leiten dabei die Untersuchung:

1. Wie werden die Daten des Korpus *RIDGES Herbology 9.0* momentan semantisch und technisch repräsentiert und welche Datenbanksysteme bilden dafür das Fundament?

2. Welche Anforderungen müsste eine mögliche Lösung erfüllen, damit diese Daten interdisziplinär nachgenutzt und weitergehend angereichert werden können?

3. Inwieweit wird die im Rahmen dieser Arbeit entwickelte Datenpipeline sowie die technische und semantische Repräsentation der Daten auf dem Fundament *Wikibase* diesen Anforderungen gerecht?

Im ersten Abschnitt dieser Arbeit werde ich die aktuelle inhaltliche Modellierung und technische Repräsentation der korpuslinguistischen Forschungsdaten im Rahmen der Architektur *ANNIS* und des Datenbanksystems *graphANNIS* vorstellen und einen Anforderungskatalog an eine darauf aufbauende Lösung zur Nachnutzung und Anreicherung entwickeln. Daran anschließend stelle ich im zweiten Abschnitt die Voraussetzungen für eine mögliche Lösung zur Nachnutzung und Anreicherung vor; dabei geht es um die Entwicklung eines eigenen Datenmodells, um *Wikibase* als Datenbanksystem im Hintergrund sowie um die nötige Datenaufbereitung. Gefolgt wird dies im dritten Abschnitt von der Präsentation und Erklärung der konkreten Datenpipeline sowie dem Vorführen exemplarischer Abfrage-/Auswertungsmöglichkeiten, die im Rahmen des neuen Datenmodells und *Wikibase* möglich werden. Den Schluss bildet eine Diskussion der vorgestellten Lösung im Hinblick auf die entwickelten Anforderungen an diese, einen Erfahrungsbericht über die Arbeit mit *Wikibase* sowie ein Ausblick auf weitere Potenziale und Herausforderungen.

Literatur

Beretta, Francesco. 2021. A challenge for historical research: Making data FAIR using a collaborative ontology management environment (OntoME). *Semantic Web* 12(2). 279–294. doi:10.3233/SW-200416. https://content.iospress.com/articles/semantic-web/sw200416. Publisher: IOS Press.

Flanders, Julia & Fotis Jannidis (eds.). 2019. *The Shape of Data in the Digital Humanities: Modeling texts and text-based resources* Digital research in the arts and humanities. London and New York: Routledge. doi:10.4324/9781315552941.

Harrower, Natalie, Maciej Maryl, Timea Biro & Beat Immenhauser. 2020. Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities. doi:10.7486/DRI.TQ582C863. Backup Publisher: ALLEA Working Group E-Humanities. https://doi.org/10.7486/DRI.tq582c863.

Kindling, Maxi & Peter Schirmbacher. 2013. Die digitale Forschungswelt als Gegenstand der Forschung / Research on Digital Research / Recherche dans la domaine de la recherche numérique. *Information - Wissenschaft & Praxis* 64(2-3). 127–136. doi:10.1515/iwp-2013-0017.

Odebrecht, Carolin, Gohar Schnelle, Catharina Fischer & Laura Perlitz. 2020. Dokumentation und Annotationsrichtlinien für das Korpus Ridges Herbology Version 9.0 (ANNIS-und PAULA-Format) auf Grundlage des Metadatenframeworks nach LAUDATIO. Stand 31.03.2020. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download-files/pubs/ridgesv9-2020-03.pdf.

Queckbörner, Boris. 2019. *Forschungsdaten und Forschungsdatenmanagement in der Geschichtswissenschaft: Gegenwärtige Praxis und Perspektiven am Beispiel ausgewählter Sonderforschungsbereiche.* Berlin: Humboldt-Universität zu Berlin Masterarbeit. doi:10.18452/20460. https://edoc.hu-berlin.de/handle/18452/21227.

Senat der Deutschen Forschungsgemeinschaft. 2015. Leitlinien zum Umgang mit Forschungsdaten. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/leitlinien_forschungsdaten.pdf.

Thalheim, Bernhard & Ivor Nissen. 2015. The Notion of a Model. In Bernhard Thalheim & Ivor Nissen (eds.), *Wissenschaft & Kunst der Modellierung. Kieler Zugang zur Definition, Nutzung und Zukunft* (Philosophische analyse = Philosophical analysis Band 64), 615–618. Boston: De Gruyter 1st edn.