Dr. David Krassnig
d.krassnig@ub.uni-frankfurt.de

Exposé für eine Masterarbeit zum Thema

Publikationspraktiken für Forschungsdaten in Hochschulschriften

Eine Untersuchung der Veröffentlichungsformate und -methoden

1 Einführung

Es gibt drei Publikationsformen für Forschungsdaten (FD) in Hochschulschriften (HSS): **(i)** vollständig in HSS integrierte Daten (z.B. Tabellen und Grafiken, die in der PDF-Datei der HSS eingebettet worden sind), **(ii)** HSS-beigefügte Daten (z.B. Dateien, die gemeinsam mit der PDF-Datei der Hochschulschrift auf den Publikationsserver der Hochschule hochgeladen worden sind) und **(iii)** auf ein separates Repositorium hochgeladene Daten, auf die innerhalb der HSS verwiesen wird [3, S. 5f.].

Im wissenschaftlichen Kontext geben präskriptive Artikel aus dem DFG-Förderprojekt „eDissPlus“ [2, 4, 5] sowie die „Policy für dissertationsbezogene Forschungsdaten“ der Deutschen Nationalbibliothek [1] vermehrt Richtlinien für den Umgang mit FD für HSS. Es fehlen bisher allerdings umfassende Studien zur Wirksamkeit bzw. Durchsetzung dieser Richtlinien bei Studierenden (z.B. durch entsprechende Prüfungsordnungen und Beratungen zu diesem Thema durch Universitätsbibliotheken). Hier existieren bisher höchstens hochspezialisierte und fachbezogene Untersuchungen.

Diese Masterarbeit beabsichtigt, hierzu eine allgemeinere Untersuchung darzubieten.

2 Forschungsfrage

Die Hauptforschungsfrage der Masterarbeit lautet „Auf welche Art und Weise wurden im institutionellen Repositorium der Leibniz Universität Hannover (LUH-Repositorium) FD von HSS bis inkl. Dezember 2023 publiziert?“ und lässt sich in folgende untergeordnete Forschungsfragen aufgliedern: **(i)** Für welchen Anteil an HSS wurden FD als Teil der PDF-Datei publiziert? **(ii)** Für welchen Anteil an HSS wurden FD als separate Datei in Form eines Supplements publiziert? **(iii)** Für welchen Anteil an HSS wurden FD in einem separaten Repositorium publiziert? **(iv)** Wie werden FD in HSS ausgezeichnet und mit dem Text der HSS verlinkt? **(v)** Wie wird in den Metadaten von HSS sichtbar gemacht, dass es zugehörige Forschungsdaten gibt?

Zusätzlich stellt sich folgende Forschungsfrage: „Inwiefern wurden Empfehlung bezüglich FD in HSS bereits in Prüfungsordnungen und anderen leitführenden Dokumenten an deutschen Universitäten verankert?“

3 Methodologie

Für die Beantwortung dieser Forschungsfragen wird der Arbeitsprozess für die Masterarbeit in vier Module aufgegliedert: **(i)** die Analyse von deutschen Promotionsordnungen und übergreifenden Richtlinien in Bezug auf FD, **(ii)** die manuelle Klassifikation der HSS im LUH-Repositorium in Bezug auf FD, **(iii)** die Auswertung der Ergebnisse aus den beiden vorherigen Modulen mit Schwerpunkt auf mögliche Handlungsempfehlungen in Bezug auf FD und **(iv)** das Training eines Modells zur automatischen Klassifizierung von HSS in Bezug auf FD auf Basis der Ergebnisse der vorhergegangen manuellen Klassifikationsarbeit.

3.1 Promotionsordnungen

Hier werden die Promotionsordnungen und andere relevante leitführende Dokumente einer einfachen Stichprobe ($n = 173$) aller promotionsberechtigter Hochschulen in Deutschland ($n = 313$) untersucht. Die Stichprobengröße wurde mit einem Konfidenzintervall von 95% und einer Fehlerspanne von 5% berechnet.

3.2 Manuelle HSS-Klassifikation

Hier wird eine mehrschichtige Stichprobe der HSS im LUH-Repositorium manuell danach klassifiziert ob die HSS, **(i)** keine FD, **(ii)** FD als Teil der PDF-Datei, **(iii)** FD als beigefügte Datei(en) oder **(iv)** FD in einem externen Repositorium haben. Die Stichprobe ist geschichtet nach den Fakultäten der LUH und nach vier 3-Jahres-Etappen. Für dieses Modul erhalte ich administrativen Zugriff auf das LUH-Repositorium. Die genaue Stichprobengröße kann erst mit diesem Zugriff berechnet werden. Die Klassifikation selber beachtet den Inhalt der PDF-Datei sowie der sich im LUH-Repositorium befindenden assoziierten Metadaten.[1]

3.3 Auswertung & Empfehlungen

Hier werden die Ergebnisse der ersten beiden Module ausgewertet und anhand der gewonnenen Daten Konzepte entwickelt, wie ein besserer Umgang mit FD in HSS erzielt werden kann und an welche Zielgruppen diese Bemühungen sich am ehesten richten sollten.

3.4 Training des Klassifikationsmodells

Hier werden die Ergebnisse der vorangegangen Klassifikationsarbeit genutzt um ein Modell zu trainieren, welches dann die restlichen HSS im LUH-Repositorium nach FD-Status klassifizieren können soll. Das Training und der Aufbau des Modells orientiert sich, zumindest erwartungsgemäß, nach Younes und Scherps Arbeit zur Identifizierung und Extraktion von Datensätzen in wissenschaftlichen Artikeln [6].

Je nachdem ob die LUH die Ressourcen für eine Kontrolle der Ergebnisse hat, wird hier entweder ein einstufiges Verfahren (direkte Identifizierung und Extraktion via ein prätrainiertes Sprachmodell wie DeBERTa in Frage-Antwort-Modus) oder ein zweistufiges Verfahren (Filterung via ein MLP mit anschließender Extraktion via ein prätrainiertes Sprachmodell wie RoBERTa) genutzt. Ersteres hat (nach bisherigen Erwartungen) eine höhere Präzision und bedarf daher weniger Nachbearbeitung, besitzt dafür aber einen geringeren Recall. Letzteres hat (nach bisherigen Erwartungen) einen höheren Recall aber dafür eine geringere Präzision.

Literatur

[1] Deutsche Nationalbibliothek [DNB]. *Policy der Deutschen Nationalbibliothek für dissertationsbezogene Forschungsdaten*. 2017. URL: https://d-nb.info/114060242X/34.

[2] Michael Kleineberg und Ben Kaden. „Zur Veröffentlichung dissertationsbezogener Forschungsdaten: Perspektiven und Kompetenzen von Promovierenden an Berliner Universitäten“. In: *Bausteine Forschungsdatenmanagement* 1 (Okt. 2018), S. 64–69. DOI: 10.17192/bfdm.2018.1.7938. URL: https://bausteine-fdm.de/article/view/7938.

[3] Susan Reilly u. a. *Opportunities of Data Exchange: Report on Integration of Data and Publications*. 2011. URL: https://hdl.handle.net/10013/epic.40198.d001.

[4] Dirk Weisbrod. „Pflichtablieferung von Dissertationen mit Forschungsdaten an die DNB – Anlagerungsformen und Datenmodell“. In: *o-bib. Das offene Bibliotheksjournal* 5.2 (Juli 2018), S. 72–78. DOI: 10.5282/o-bib/2018H2S72-78.

[5] Dirk Weisbrod, Ben Kaden und Michael Kleineberg. „eDissPlus – Optionen für die Langzeitarchivierung dissertationsbezogener Forschungsdaten aus Sicht von Bibliotheken und Forschenden“. In: *E-Science-Tage: Forschungsdaten managen*. Humboldt-Universität zu Berlin, 2017, S. 189–198. DOI: 10.18452/22310.

[6] Yousef Younes und Ansgar Scherp. „Question Answering Versus Named Entity Recognition for Extracting Unknown Datasets“. In: *IEEE Access* 11 (2023), S. 92775–92787. DOI: 10.1109/ACCESS.2023.3309148.

[1] Hier sollte angemerkt werden, dass die Metadaten noch nicht explizit anmerken, wenn FD in HSS vorhanden sind. Dies soll sich, auch im Rahmen dieser Masterarbeit, in Zukunft verändern und bei Aufnahme einer HSS erfasst werden.