Analyse, Produktion, Reflexion: Nachnutzungsszenarien für Forschungsdaten am Beispiel der Daten des Projekts Dehmel digital

Bläß, Sandra
sandra.blaess@uni-hamburg.de
Universität Hamburg, Deutschland

Flüh, Marie
marie.flueh@uni-hamburg.de
Universität Hamburg, Deutschland

Nantke, Julia
julia.nantke@uni-hamburg.de
Universität Hamburg, Deutschland

Reul, Christian
christian.reul@uni-wuerzburg.de
Universität Würzburg, Deutschland

Das Ziel wissenschaftlicher Editionen besteht seit jeher in der Nachnutzung durch die (wissenschaftliche) Community. Unter den Vorzeichen von Open Data und Open Science ändern sich allerdings die Möglichkeiten der Bereitstellung und Nachnutzung des erschlossenen Materials. Gleichzeitig haben Wissenschaftler:innen, die mit Methoden der Digital Humanities arbeiten, andere Anforderungen und Bedarfe an bereitgestellte Daten z.B. im Hinblick auf den Umfang der Korpora und die spezifischen Datentypen. Ziel unseres Beitrags ist es, anhand der unterschiedlichen, von uns im digitalen Editions- und Forschungsprojekt *Dehmel digital* (Nantke 2022) produzierten Datentypen systematisch Szenarien der digitalen Nachnutzung durchzuspielen und anhand von Beispielen zu präsentieren. Die Darstellung ist projektbezogen und nicht erschöpfend in Bezug auf alle denkbaren Datentypen und Nutzungsmöglichkeiten. Dennoch sollen die dargestellten Szenarien in ihrer Bandbreite exemplarisch auch für andere Projektkontexte fungieren können.

Wir beziehen uns auf folgende Datentypen: 1) Metadaten von Briefen unterschiedlicher Schreibender aus dem Korrespondenznetz von Ida und Richard Dehmel, 2) digitale Bilder der Dokumente, 3) maschinenlesbarer Text der Briefe, 4) Annotationen von Entitäten sowie 5) algorithmische Modelle.

In Abhängigkeit vom jeweiligen Datentyp ergeben sich unterschiedliche Nachnutzungsszenarien. Diese reichen von dem aus editorischer Sicht klassischen Szenario der Nutzung der bereitgestellten Daten in (literatur)wissenschaftlichen Analysen über die Nutzung zur Produktion eigener Korpora und Modelle, die dann wiederum Gegenstand der Nachnutzung werden können, bis hin zur Algorithmen-gestützten Reflexion der konzeptuellen Grundlagen einer solchen Datensammlung.

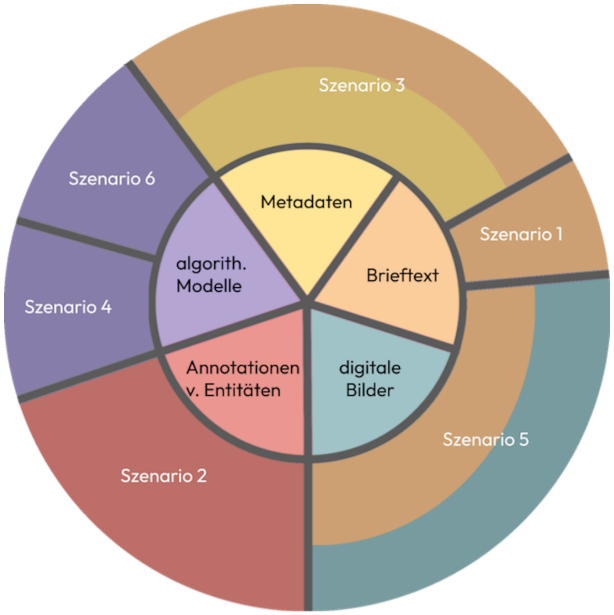

Abb.1: Datentypen und Nachnutzungsszenarien im Überblick

Szenario 1: Analyse von Briefinhalten auf der Basis von Datentyp 3

Briefe stellen relevante Quellen für die Rekonstruktion historischer Diskurse dar (Baillot 2011). Diese in einem Gesamtüberblick und nicht nur in Einzelbeispielen zu erfassen, ist mittels Close Reading-Verfahren kaum zu bewältigen. Ein zentrales computergestütztes Nachnutzungsszenario für unsere Daten sind daher Analysen mittels Distant Reading-Verfahren. Auf der Basis der erzeugten Transkripte kann u.a. eine automatisierte Exploration der zentralen Briefinhalte über Topic Modeling umgesetzt werden (Andorfer 2017; Henny-Kramer/Neuber 2023). Ergänzend hierzu lassen sich z.B. stimmungsmäßige Gewichtungen in den Briefen durch Sentiment Analysis ermitteln und zu den Topics ins Verhältnis setzen.

Szenario 2: Korrespondenznetze sichtbar machen auf der Basis von Datentyp 4

In den im Projekt *Dehmel digital* produzierten Daten sind Entitäten (Personen, Orte, Institutionen, Werke) annotiert. Netzwerkanalysen auf der Basis dieser Annotationen bieten die Möglichkeit, Dynamiken innerhalb der vernetzten Kommunikationspraxis offenzulegen und genauere Einblicke in personelle Kontakte, organisatorische Strukturen und räumliche Bewegungen zu erlangen (Nantke/Bläß/Flüh 2022).

Szenario 3: Vernetzung mit anderen Briefeditionen auf Basis von Datentyp 1 und 3

Die von uns erzeugten Briefmetadaten können über die Plattform *correspSearch* abgerufen werden. Dadurch können Nachnutzende unsere Daten im Rahmen individueller Suchanfragen in Kombination mit den Daten anderer Briefeditionen nutzen.

Szenario 4: Texte erschließen auf der Basis von Datentyp 5

Neben den erschlossenen Dokumenten stellen wir auch die im Projekt von uns trainierten HTR- und NER-Modelle zur Nachnutzung zur Verfügung. Auf Basis dieses Datentyps können weitere Dokumente, die nicht Teil des Projekts sind, erschlossen und somit neue Daten für die weitere Nachnutzung produziert werden. Dies gilt zum einen für handschriftliche Dokumente der Schreibenden, für die wir HTR-Modelle trainiert haben (z.B. Stefan Zweig, Detlev v. Liliencron, Julie Wolfthorn). Zum anderen können die Named Entity-Classifier für die Erschließung weiterer deutschsprachiger Briefe aus einem ähnlichen Zeitraum genutzt werden. Es besteht auch die Möglichkeit, auf der Basis unserer Trainingsdaten spezifische Modelle für andere Anwendungsfälle nachzutrainieren (Flüh/Lemke 2022).

Szenario 5: gemischte HTR-Modelle trainieren auf Basis von Datentyp 2 und 3

Die in *Dehmel digital* teilautomatisiert generierten, qualitativ hochwertigen Transkripte zahlreicher unterschiedlicher Schreibender bieten in Kombination mit den zugehörigen Bilddigitalisaten den idealen Ausgangspunkt für das Training sog. 'gemischter Modelle', mit deren Hilfe sich deutlich mehr unterschiedliche Handschriften aus dem Zeitraum um 1900 transkribieren lassen.

Szenario 6: Reflexion der theoretischen Fundierung von Datensammlungen auf Basis von Datentyp 5

Unsere NER-Classifier wurden auf den Dokumententyp 'Brief um 1900' trainiert. Eine experimentelle Anwendung z.B. des Orte-Classifiers auf ein Korpus mit Texten eines deutlich abweichenden Dokumententyps (z.B. fiktionale Texte) kann insbesondere in Kombination mit einem auf den Dokumententyp zugeschnittenen Classifier dazu beitragen, die theoretisch-konzeptuelle Fundierung offenzulegen, welche in die Modellierung des Classifiers eingegangen ist, indem die Ergebnisse der Classifier vergleichend betrachtet werden (vgl. dazu die Fallstudie von Flüh/Schumacher/Nantke im Erscheinen).

Bibliographie

Andorfer, Peter. 2017.: " Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich". *Zeitschrift für digitale Geisteswissenschaften* 2017. https://doi.org/10.17175/2017_002 .

Baillot, Anne. 2011. *Netzwerke des Wissens. Das intellektuelle Berlin um 1800* . Berliner Wissenschafts-Verlag.

Flüh, Marie, Mareike Schumacher und Julia Nantke. Im Erscheinen. " Place and Space in Literature. Named Entity Recognition as a Possibility for Spatial Modelling in Computational Literary Studies". In: *Geography Meets Digital Humanities* , hg. von Finn Dammann und Dominik Kremer. Bielefeld: transcript.

Flüh, Marie, und Marc Lemke. 2022. " An Experimental Attempt to Use Transfer Learning for Named Entity Recognition in Letters from the 19th and 20th Century". *DH2022* , 2022.

Henny-Krahmer, Ulrike und Frederike Neuber. 2023. " Topic Modeling in Digital Scholarly Editions". *Machine Learning and Data Mining for Digital Scholarly Editions* , hg. von Bernhard Geiger u. a., Bd. 18, Books on Demand.

Nantke, Julia. 2022. *Dehmel digital.* hg von ders. unter Mitarbeit von Sandra Bläß und Marie Flüh. https://dehmel-digital.de [zugegriffen: 21. Juli 2022].

Nantke, Julia, Sandra Bläß, Marie Flüh und David Maus. 2022. " Best of Both Worlds. Zur Kombination algorithmischer und manueller Verfahren bei der Erschließung großer Handschriftenkorpora". *DHd 2022. Konferenzabstracts* , 2022, https://doi.org/10.5281/zenodo.6328113 .