Randnotizen des produktiv Unsicheren: Über die Dokumentation von Sackgassen

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg, Deutschland
ORCID: 0000-0002-9548-8461

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg, Deutschland
ORCID: 0000-0003-1016-2911

Helling, Patrick
patrick.helling@uni-koeln.de
Institut für Deutsche Philologie, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Universität Würzburg, Deutschland
ORCID: 0000-0003-4043-165X

Gehen wir die Wege des produktiv Unsicheren, hinterlassen wir im Idealfall Spuren. Diese können in Form von beschriebenen Ergebnissen, erzeugten Ressourcen und Dokumentation deutlich werden. Erweist sich eine Untersuchung als erfolgreich oder entsteht eine nachnutzbare Ressource werden die Ergebnisse in der Regel mittels Veröffentlichungen und / oder Metadaten beschrieben.

Was aber bleibt, wenn ein Ansatz in einer Sackgasse endet, also z. B. eine Untersuchung ohne Ergebnis bleibt oder keine Ressource erzeugt werden kann? Liegen keine groben methodischen Fehler vor, ergibt sich zumindest ein Erkenntnisgewinn, der aber selten zugänglich gemacht werden kann.[1] Quoten für Negativresultate in Zeitschriften und bei Konferenzen (Brazil 2024) oder die Präregistrierung von Studien (Neuendorf und Rüdiger 2021) sind mögliche Wege dem Erkenntnisgewinn aus Sackgassen Platz einzuräumen. Ebenso wird von der DFG die zitierbare Darlegung von Negativresultaten in Abschlussberichten vorangetrieben (Deutsche Forschungsgemeinschaft 2023).

Nichtsdestoweniger sind die Hürden einer adäquaten Beschreibung von Sackgassen ungleich höher. Selten markiert die Sackgasse das Ende des Forschungsvorhabens, wer kann es sich also leisten, ausreichend Zeit in eine ausführliche Dokumentation der Sackgasse zu investieren, statt die Arbeiten in eine andere Richtung fortzusetzen? Denn gleichzeitig ist die Beschreibung der Sackgasse nur dann nützlich, wenn sich zukünftige Vorhaben mit dem Vorgehen, das in die Sackgasse führte, ausreichend vergleichen lassen, um damit die Sackgasse als daten- oder werkzeuggebunden zu identifizieren oder deren Aussage über grundlegende methodische Zusammenhänge zu bestätigen. Dafür müssen aber nicht nur das grobe Vorgehen, sondern alle Verarbeitungsschritte im Detail beschrieben werden: für automatische Verarbeitungsschritte können das z. B. Werkzeuge, Wissensbasen, Versionen, Parameter etc. sein, für manuelle Schritte z. B. Richtlinien, Kriterien, theoretische Grundlagen, Trainingsstatus von Annotierenden etc. Auch Entscheidungen, die ggf. aus pragmatischen Gründen getroffen werden und meist nicht in Publikationen auftauchen, wie zum Beispiel das (ggf. unsystematische) Verbessern von Rechtschreibfehlern oder die Beschränkung einer Segmentgröße zur besseren automatischen Verarbeitbarkeit können Einfluss auf die Vergleichbarkeit der Arbeitsabläufe haben.

Forschungsbegleitende Prozessmetadaten

Um Arbeitsabläufe und automatische wie manuelle Verarbeitungsschritte zu dokumentieren, eignen sich Prozessmetadaten. In Jung (2020) stellen wir aus der Sicht der Reproduzierbarkeit und Nachnutzbarkeit (erfolgreicher) Arbeitsabläufe folgende Aspekte als relevant für die Sammlung von Prozessmetadaten, u. A. in den Computational Literary Studies, vor:

- Auf Grundlage welcher Daten (+ Version) wird der Schritt durchgeführt?
- Welche Operation wird auf den Ausgangsdaten durchgeführt? (Analyse, Selektion, Interpretation, ...; manuell, automatisch, semi-automatisch)
- Wer oder was führt die Operation aus?
- Person(en): Name / Kürzel / ID (Datenschutz bedenken)
- Werkzeug(e): Name / ID, Version, Einstellung und Parameter, ggf. Betriebssystem / Hardware
- genutzte Komponenten (Annotationsrichtlinien, Lexikon, Modell, etc.): Name / ID, Version
- ggf. Zusammenhänge zwischen Ressourcen (Korpus + Version auf dem das Modell trainiert wurde, ...)
- ggf. neu entstandene Daten (Subkorpus, Annotationsebene, ...): Name / ID, Version

Auch für die Dokumentation von Sackgassen eignen sich diese Metadaten, da sie die Vergleichbarkeit von Arbeitsabläufen ermöglichen. Des weiteren erübrigt sich die Frage, ob es sich lohnt, Zeit für die Dokumentation von Sackgassen aufzuwenden, wenn Prozessmetadaten von Anfang an systematisch mitgeschrieben werden. Im Falle einer Sackgasse können ohne zusätzlichen Aufwand deren Randnotizen, die Prozessmetadaten, als Dokumentation zur Verfügung gestellt werden, während sie auch für erfolgreich abgeschlossene Vorhaben die Nachnutzbarkeit erhö-

hen. Dabei sind Prozessmetadaten durchaus zusätzlich zu ggf. etablierten elaborierten Dokumentationsformen zu sehen.

Während Formate wie der TEI Header (The TEI Consortium 2024) oder PROV (Moreau u. a. 2015) Prozessinformationen beschreiben können, liegt ihnen die Idee eines Ergebnisses zugrunde, dessen Erstellungsweg beschrieben wird. Prozessmetadatenschemata, die den Fokus auf die Abläufe legen, wie z.B. RePlay-DH (Gärtner, Hahn und Hermann 2018) sind daher näher am Verständnis der Beschreibung einer Sackgasse.

Eine Galerie des gut dokumentierten Scheiterns?

Gehen wir also einmal kurz davon aus, dass wir uns im Fachbereich auf ein (oder mehrere abbildbare) Formate zur Prozessdokumentation einigen könnten, diese in fachspezifische FDM-Schulungen und -Dokumentation aufnähmen und das kontinuierliche (ggf. teil-automatische) Mitschreiben von Prozessmetadaten in unsere tägliche Routine übernähmen, wie machen wir den Erkenntnisgewinn der Community zugänglich? Können wir uns eine Datenbank der Sackgassen vorstellen, zu der wir mit unseren Prozessbeschreibungen beitragen? Denkbar wäre hier eine Sammlung wie z.B. die der Workflows im SSH Open Marketplace (Concordia, Meghini und Benedetti 2020) oder auch die von Jablonka, Patiny und Smith (2022) vorgeschlagene Kombination von Positiv- und Negativresultaten, wobei aufgrund der Gleichwertigkeit von manuellen und automatischen Schritten in den Workflows der DH, das Ziel nicht die Erstellung von 'machine actionable' Daten sein kann.

Wären wir bereit, eine solche Datenbank vor Beginn unserer Vorhaben zu durchsuchen? Wie hoch wäre tatsächlich die Vergleichbarkeit einzelner Ketten von Verarbeitungsschritten? Diese Fragen möchten wir mit den Teilnehmenden der DHd Jahrestagung am Poster diskutieren.

Sicher kann es nicht das Ziel sein, durch einzelne Sackgassen Methoden, Datensätze oder Werkzeuge aus unseren Vorhaben zu verbannen. Im Gegenteil sollten wir uns aufgrund der Sackgassen die Fragen stellen, warum bestimmte Konstellationen gescheitert sind und ggf. das mögliche Scheitern durch Veränderung einzelner Komponenten in Frage stellen, bevor wir eine Methode komplett verwerfen. Eine entsprechende Menge an Ketten von Verarbeitungsschritten könnte aber durchaus eine systematische Auswertung, ggf. auch über die Grenzen der Teilbereiche der Digital Humanities hinweg ermöglichen und damit konkrete neue Vorhaben schärfen oder sogar neue Fragestellungen aufwerfen.

Fußnoten

1. Dabei können sowohl die methodischen Fehler als auch der Erkenntnisgewinn aus verschiedenen der von Gengnagel (2022) aufgeschlüsselten Problemfeldern stammen: Versagen von Technologien, menschliches, arbeitspraktisches sowie intellektuelles Versagen.

Bibliographie

Brazil, Rachel. 2024. „Illuminating 'the ugly side of science ': fresh incentives for reporting negative results.“ 8. Mai. https://www.nature.com/articles/d41586-024-01389-7 (zugegriffen: 24. Juli 2024).

Concordia, Cesare, Carlo Meghini a nd Filippo Benedetti. 2020. „Store scientific workflows data in SSHOC repository.“ In *Proceedings of the workshop about language resources for the SSH cloud*, hg. von Daan Broeder, Maria Eskevich, und Monica Monachini, 1–4. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lr4sshoc-1.1.

Deutsche Forschungsgemeinschaft, Hrsg. 2023. „Deutsche Forschungsgemeinschaft schafft Grundlagen für die Veröffentlichung von Abschlussberichten.“ 2. Januar. https://www.dfg.de/de/aktuelles/neuigkeiten-themen/info-wissenschaft/2023/info-wissenschaft-23-01 (zugegriffen: 24. Juli 2024).

Gärtner, Markus, Uli Hahn und Sibylle Hermann. 2018. „Supporting sustainable process documentation“. In *Language technologies for the challenges of the digital age*, hg. von Georg Rehm und Thierry Declerck, 284–291. Cham: Springer International Publishing. doi:10.1007/978-3-319-73706-5_24.

Gengnagel, Tessa. 2022. „Vom Topos des Scheiterns als konstituierender Kraft: Ein Essay über Erkenntnisprozesse in den Digital Humanities.“ In *Fabrikation von Erkenntnis – Experimente in den Digital Humanities (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5),* hg. von Manuel Burghardt, Lisa Dieckmann, Steyer Timo, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis, Ulrike Wuttke. Wolfenbüttel 2022. Version 2.0 vom 19.09.2024. HTML / XML / PDF. DOI: 10.17175/sb005_011_v2

Jablonka, Kevin Maik, Luc Patiny und Berend Smith. 2022. „Making the collective knowledge of chemistry open and machine actionable“. *Nature chemistry* 14: 365–376. doi:10.1038/s41557-022-00910-7, .

Jung, Kerstin. 2020. „FAIR in der Praxis - erster Beitrag“. 22. Dezember. https://dfg-spp-cls.github.io/de/2020/12/22/FAIR-I.html (zugegriffen: 24. Juli 2024).

Moreau, Luc, Paul Groth, James Cheney, Timothy Lebo und Simon Miles. 2015. „The rationale of PROV“. *Journal of web semantics* 35: 235–257. doi:10.1016/j.websem.2015.04.001.

Neuendorf, Claudia und Christin Rüdiger. 2021. „Präregistrierung von Studien in der empirischen Bildungsforschung - Wozu, Wie und Wo?“ *Workshop-*

reader digiGEBF open science summer 2021. https://osf.io/2hdxn/.

The TEI Consortium. 2024. „The TEI Header“. In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 4.8.0. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html.