Löschen, Behalten, Ersetzen und Vertauschen: Open Science dank abgeleiteter Textformate

Philippe Genêt[1], **Lukas Weimer**[2]
[1]Deutsche Nationalbibliothek, [2]Niedersächsische Staats- und Universitätsbibliothek Göttingen

Als eines von 26 Konsortien der Nationalen Forschungsdateninfrastruktur (NFDI) hat sich Text+ zum Ziel gesetzt, sprach- und textbasierte Forschungsdaten langfristig zu erhalten, verfügbar und für die Forschung nutzbar zu machen. Dabei basiert Text+ auf einer verteilten Infrastruktur von Daten- und Kompetenzzentren, die ihre Forschungsdaten, ihre Dienste und ihre Expertise über Schnittstellen in das Netzwerk von Text+ einbringen. Daten und Dienste sind dadurch zentral abfrag- und bearbeitbar, bspw. über die Text+ Registry oder die Federated Content Search, die Metadaten- und Volltextsuche in angebundenen Ressourcen ermöglichen. Zusammengeführt und zentral zugänglich sind sie über das Text+ Portal.

Während viele der Sprach- und Textressourcen aus den Datenzentren von Text+ der Wissenschaft frei zur Verfügung stehen, sind andere wegen rechtlicher Einschränkungen nur bedingt für die Forschung nutzbar. Zu nennen sind hier insbesondere Werke, die urheberrechtlich (zum Beispiel weite Teile der Sammlung der Deutschen Nationalbibliothek), oder durch Daten- und Persönlichkeitsrechte geschützt sind (häufig in Interviews wie denen des Hamburger Zentrums für Sprachkorpora), bzw. lizenzrechtlichen Limitierungen unterstehen und etwa nur im Rahmen einer Lizenzvereinbarung zugänglich sind (wie etwa das Wortnetz GermaNet des Seminars für Sprachwissenschaft an der Universität Tübingen).

Für Text+ liegt die Herausforderung darin, auch derart geschützte Werke – im Sinne der Open Science – für die Wissenschaft zugänglich und verwertbar zu machen. Eine Möglichkeit dazu bieten abgeleitete Textformate (ATF)[1].

Das Ziel von ATF ist es, die zuvor für die spätere Analyse mit Annotationen angereicherten Textinformationen selektiv durch Behalten, Löschen, Ersetzen und Vertauschen so stark zu reduzieren, dass das Ergebnis keinen rechtlichen Beschränkungen mehr unterliegt, die noch vorhandenen Informationen jedoch für bestimmte Aufgaben des Text- und Data-Minings (TDM) ausreichen. Insbesondere im Hinblick auf das Urheberrecht muss dabei besonders darauf geachtet werden, dass ATF weder die Wiedererkennbarkeit, den Werkgenuss noch die Wiederherstellbarkeit des Originaltextes ermöglichen. Dennoch bleiben ATF für gängige Textanalysemethoden auswertbar, wie aktuelle Untersuchungen[2] zeigen.

---

[1] vgl. Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, Jörg Röpke: Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel 2020. text/html Format. DOI: 10.17175/2020_006

[2] vgl. Keli Du: Understanding the impact of three derived text formats on authorship classification with Delta. DHd 2023 Open Humanities Open Culture. 9. Tagung des Verbands “Digital Humanities im deutschsprachigen Raum” (DHd 2023), Trier, Luxemburg, 10. März 2023. https://doi.org/10.5281/zenodo.7715299.

Die Herstellung von ATF basiert auf Methoden des TDM und fußt daher rechtlich auf Grundlage der 2021 eingeführten sogenannten TDM-Schranke[3] im deutschen Urheberrechtsgesetz. Sie erlaubt das TDM für Zwecke der nichtkommerziellen wissenschaftlichen Forschung – auch durch Kulturerbeeinrichtungen wie Bibliotheken und Archive.

Text+ arbeitet an einem Vorschlag für einen DIN-Standard zu ATF sowie an einer Veröffentlichung zu rechtlichen Aspekten dieser Formate, um wissenschaftliche Communitys, die diese Daten nutzen, und Institutionen, die solche Daten bereitstellen möchten, gleichermaßen mit der nötigen Expertise zu unterstützen. Darüber hinaus planen an Text+ beteiligte Institutionen nicht nur die Veröffentlichung bestimmter ATF, sie bieten zugleich auch nützliche Tools und Services für den Umgang mit diesen Formaten an. So stellt bspw. die SUB Göttingen die Natural Language Processing-Pipeline MONAPipe zur Verfügung, die nicht nur zahlreiche Module für die Verarbeitung, sondern perspektivisch auch für die Herstellung von ATF enthalten wird. Auch das von der Universität Tübingen bereitgestellte Annotationstool WebLicht gehört zu diesen Angeboten.

Unser Poster zeigt, wie Text+ durch ATF auch urheberrechtlich geschützte Werke für die wissenschaftliche Forschung zur Verfügung stellt und diese in die eigene Infrastruktur einbettet, die neben Zugänglichkeit und Auffindbarkeit auch Dienste zur weiteren Verarbeitung bietet. So stellen wir ATF als Instrument vor, um offene Wissenschaft zu befördern und die Bearbeitung rechtebewehrter textueller Ressourcen zu ermöglichen.

Die Autoren:

Philippe Genêt (https://orcid.org/0009-0001-5095-8052) ist Theater-, Film- und Medienwissenschaftler und Amerikanist. Er arbeitet in der Deutschen Nationalbibliothek (DNB) und koordiniert dort alle Aktivitäten der DNB in der NFDI. Im Konsortium Text+ koordiniert er die Datendomäne Collections.

Dr. Lukas Weimer (https://orcid.org/0000-0001-6919-3646) arbeitet in der Abteilung Forschung und Entwicklung der Niedersächsischen Staats- und Universitätsbibliothek Göttingen im Office des NFDI-Konsortiums Text+ sowie im Office des Verbunds Base4NFDI. Seinen Hintergrund hat er in der neueren deutschen Literaturgeschichte.

[3] § 60d UrhG, vgl.: https://www.gesetze-im-internet.de/urhg/__60d.html