Auszugsweise Veröffentlichung geschützter Werke mit Annotationen steuern

Gärtner, Markus
markus.gaertner[at]ims.uni-stuttgart.de
Universität Stuttgart, Deutschland
ORCID-iD: 0000-0002-2687-4350

Kleinkopf, Felicitas
felicitas.kleinkopf[at]kit.edu
Karlsruher Institut für Technologie (KIT), Deutschland
ORCID-iD: 0000-0001-8670-2668

Andresen, Melanie
melanie.andresen[at]ims.uni-stuttgart.de
Universität Stuttgart, Deutschland
ORCID-iD: 0000-0002-3913-1273

Hermann, Sibylle
sibylle.hermann[at]ub.uni-stuttgart.de
Universität Stuttgart, Deutschland
ORCID-iD: 0000-0001-9239-8789

Zusammenfassung. Die Herausgabe oder Veröffentlichung zugrundeliegender Daten zum Zwecke der Evaluation oder Nachnutzung durch Dritte ist ein wesentlicher Bestandteil der guten wissenschaftlichen Praxis. Dem stehen diverse Einschränkungen des geltenden Urheberrechts entgegen, wodurch insbesondere an literarischen Werken interessierte Forschungsdisziplinen massiv beeinträchtigt werden, da nur auf gemeinfreien Texten basierende Korpora oder andere Forschungsdaten uneingeschränkt weitergegeben werden können. Zwar existieren Möglichkeiten, geschützte Ursprungsdaten derart zu verändern und weiterzuverwenden, dass die endgültige Form nicht mehr dem ursprünglichen Urheberrecht unterliegt, allerdings ist dies für viele Anwendungsfälle keine zufriedenstellende Lösung. Wir stellen als Alternative das XSample Konzept vor, welches auf urheberrechtlichen Schranken basiert und es ermöglicht, kleine Auszüge geschützter Werke und Annotationen in Reinform zugänglich zu machen. Da selbige Auszüge durch ihren geringen Umfang die Gefahr bergen, nicht die für konkrete Interessen oder Forschungsfragen relevanten Datenpunkte zu enthalten, wird zusätzlich eine Korpusanfrageschnittstelle eingesetzt. Somit können Nutzer*innen ihre Bedarfe gezielt mittels einer formalen Anfragesprache definieren und die Nützlichkeit der

generierten Auszüge maximieren. Die prototypische Implementierung unseres Ansatzes setzt auf weit verbreitete Standards und Technologien und ist mit geringem Aufwand in existierende Infrastruktur integrierbar.

1 Problematik

Ein Großteil der Forschung in den Digitalen Geisteswissenschaften und anderen textbasierten Disziplinen sieht sich seit langer Zeit mit dem Problem konfrontiert, dass das geltende Urheberrecht die Weitergabe der gewonnenen Forschungsdaten erheblich erschwert, oder mitunter gänzlich unmöglich macht. Dies führte zu der immer noch andauernden Situation, dass Forschungsfragen in textverarbeitenden Disziplinen oftmals nicht am eigentlichen wissenschaftlichen Interesse und Bedarf ausgerichtet sind, sondern vielmehr basierend auf der Verfügbarkeit und (Weiter-) Verwendbarkeit der zugrundeliegenden oder angestrebten Textressourcen formuliert werden. Insbesondere Annotationsprojekte arbeiten seit jeher nahezu ausschließlich auf gemeinfreiem Material statt an zeitgenössischen Werken, da nur aus gemeinfreiem Inhalt aufgebaute Korpora auch veröffentlicht werden dürfen, was die gute wissenschaftliche Praxis erfordert. Insbesondere zum Zwecke späterer Nachprüfung und Evaluation für Anschlussforschungen ist eine öffentliche Verfügbarkeit erstrebenswert, wenn nicht sogar zwingend erforderlich. Zwar lassen sich individuelle Absprachen mit Rechteinhabern erzielen, um Arbeiten auf geschützten Werken zu ermöglichen und die resultierenden Forschungsdaten weitergeben zu können. Allerdings steht der damit häufig verbundene zeitliche Aufwand in direktem Widerspruch zur vergleichsweise kurzen Laufzeit typischer Forschungsprojekte. Im Folgenden wird auf Basis der aktuellen (deutschen) Rechtslage ein Konzept vorgestellt, das als Kompromiss die Interessen der Rechteinhaber schützen und gleichzeitig die auszugsweise Herausgabe von Textkorpora zur Nachnutzung ermöglichen soll.

2 Rechtliche Möglichkeiten

Neben expliziten Absprachen zwischen Rechteinhabern und Forschenden boten sich bisher nur begrenzte Möglichkeiten, um Forschung auf geschützten Texten durchzuführen und dabei nicht automatisch die Nachnutzbarkeit durch Nichtherausgabe von Forschungsdaten zu beeinträchtigen. Das kürzlich vorgestellte Konzept der abgeleiteten Textformate[1] beschreibt Verfahren, um Originaltexte geschützter Werke

[1] Schöch 2020.

durch verschiedene Methoden derart zu obfuskieren, dass die endgültige Repräsentation nicht mehr dem ursprünglichen Urheberrecht unterliegt und somit ohne Einschränkungen weitergegeben werden darf. Während dieser Ansatz für bestimmte Anwendungsfälle nutzbar ist, erfordern insbesondere Disziplinen mit einem hohen Anteil interpretativer Analysen zwingend Zugriff auf den originalen Kontext textueller Untersuchungsgegenstände. Dies gilt auch für die Evaluation von Anwendungsfällen mit ursprünglich rein quantitativen Untersuchungsmethoden.
Mittlerweile wurde die Problematik auch seitens des Gesetzgebers aufgegriffen: § 60d des Urheberrechtsgesetzes (UrhG) erlaubt seit 2018 die Nutzung urheberrechtlich geschützter Werke zum Zwecke der nichtkommerziellen Forschung mittels Text und Data Mining. Integriert wurden daneben auch Regularien zur Zugänglichmachung für Dritte während eines laufenden Projekts, womit zumindest in Veröffentlichungsverfahren innerhalb der aktiven Projektlaufzeit die Weitergabe im Kontext von Review-Prozessen ermöglicht wird. Nachdem der europäische Gesetzgeber im Jahr 2019 die sog. DSM-Richtlinie verabschiedet hat, erfolgten zum Juni 2021 rechtliche Änderungen im deutschen UrhG. Leider wurde in diesem Zuge keine ausdrückliche Möglichkeit geschaffen, wie Korpora zu Zwecken von Anschlussforschungen nachgenutzt werden können.

3 Auszugsweise Veröffentlichung

Im aktuell laufenden Projekt „XSample“ wurde unter Berücksichtigung der wissenschaftlichen Bedarfe und juristischen Rahmenbedingungen[2] ein Workflow-Konzept erarbeitet, wie bibliothekarische Einrichtungen als zentrale Akteure die Nachnutzbarkeit von geschützten Textkorpora unterstützen können. Dieses Konzept stützt sich auf die Verknüpfung der urheberrechtlichen Schranken des § 60c[3] und § 60d UrhG. Mittels einer als Webinterface implementierten Benutzerschnittstelle soll es Interessierten so auf einfachem Wege ermöglicht werden, basierend auf inhaltlichen Kriterien auszugsweise Zugang zu geschützten Korpora zu erhalten, um beispielsweise damit zusammenhängende Veröffentlichungen oder Ergebnisse zu validieren oder deren Eignung für Anschlussforschung im Kontext eigener Forschungsfragen zu testen. Eine

[2] Kleinkopf 2021.
[3] Diese urheberrechtliche Schranke erlaubt die anteilige Weitergabe von bis zu 15% geschützter Werke zum Zwecke der nichtkommerziellen Forschung.

prototypische Implementierung des Konzepts als JSF-Web-Applikation befindet sich derzeit in der Entwicklung und ist quelloffen[4] verfügbar. Abbildung 1 zeigt eine grobe Übersicht der am Workflow beteiligten Akteure und den Datenfluss von Primärdaten über erzeugte Forschungsdaten oder Annotationen bis hin zu generierten Auszügen zur Nachnutzung durch weitere Projekte. Aus Platzgründen wird hier auf eine detailliertere Beschreibung der Architektur[5] verzichtet. In der oberen Schleife erzeugen Forschende auf Basis geschützter Werke Forschungsdaten und pflegen diese in ein nicht-öffentliches Repositorium ein, bzw. stellen Metadaten in einem öffentlichen Teil des Repositoriums bereit, um die Auffindbarkeit der Ressourcen sicherzustellen. Durch den hohen Verbreitungsgrad von Repositorien-Systemen wie Dataverse[6] mit feingranularem Rechtesystem und der Anbindbarkeit an lokal existierende Authentifizierungsschnittstellen erfüllen Bibliotheken oder ähnliche Institutionen an Universitäten häufig bereits die infrastrukturellen Voraussetzungen für diesen Workflowabschnitt. Ausgehend von den öffentlich verfügbaren Metadaten können Interessierte nach ebenfalls erfolgter Authentifizierung die Generierung individueller Auszüge anstoßen.

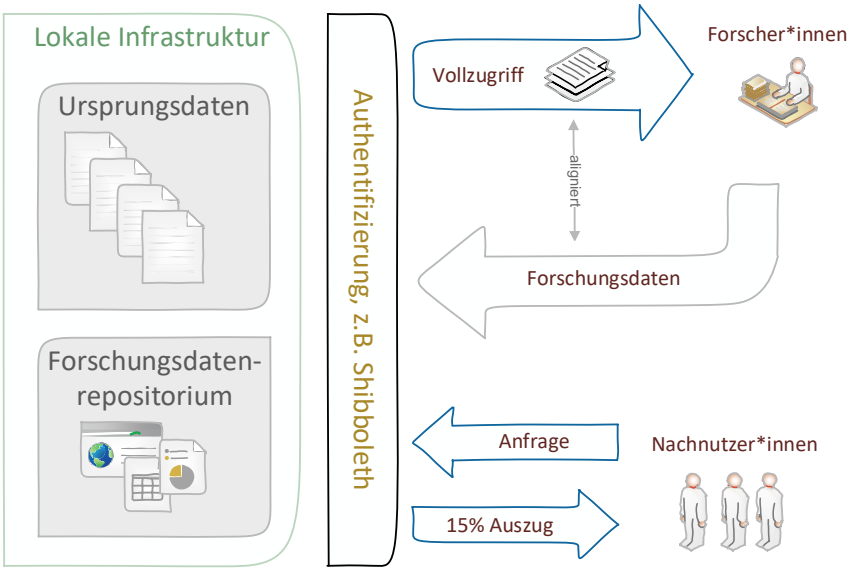

Abb. 1. Akteure und Architekturüberblick

[4] https://github.com/ICARUS-tooling/xsample-server.
[5] Gärtner 2021.
[6] The Dataverse Project, https://dataverse.org/.

4 Anfragegesteuerte Auszugerstellung

Je nach Anwendungsfall oder Forschungsfrage kann es sich als wenig zielführend herausstellen, wenn eine statische (z.B. die ersten 15%) oder zufällig zusammengestellte Teilmenge eines Textkorpus als Auszug herausgegeben wird. Um die Nützlichkeit der im Auszug enthaltenen Daten für die Nachnutzung zu maximieren, wird bei der Anfrage zur Auszuggenerierung ermöglicht, eine integrierte Korpusanfrageschnittstelle[7] zu nutzen. Anhand der im Korpus enthaltenen Annotationen kann das Interesse an bestimmten Phänomenen formal spezifiziert werden, beispielsweise an bestimmten linguistischen, aber auch allen anderen in Annotationen hinterlegten Merkmalen. Zusätzlich stehen verschiedene weitere Optionen zur Verfügung, um den Inhalt erzeugter Auszüge optimal auf die individuellen Bedürfnisse anzupassen, ohne jedoch voreilig Zugriff auf die geschützten Daten an sich zu gewähren. Auf diese Weise kann sichergestellt werden, dass Nutzer*innen im Idealfall genau den Anteil als Auszug erhalten, der für sie am interessantesten ist und sie somit daraus den größtmöglichen Nutzen unter aktuell geltenden Einschränkungen ziehen können.

Bibliografie

Gärtner, Markus. 2020. “The Corpus Query Middleware of Tomorrow – A Proposal for a Hybrid Corpus Query Architecture“, *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*: 31-39. https://www.aclweb.org/anthology/2020.cmlc-1.5

Markus Gärtner, Felicitas Kleinkopf, Melanie Andresen and Sibylle Hermann: "Corpus Reusability and Copyright - Challenges and Opportunities", *Proceedings of the 9th Workshop on Challenges in the Management of Large Corpora*: S.10-19. https://doi.org/10.14618/ids-pub-10467.

Kleinkopf, Felicitas, Janina Jacke, Markus Gärtner. 2021. “Text- und Data-Mining“, *MMR 2021*: 196-200.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020. “Abgeleitete Textformate:

[7] Gärtner 2020.

Text und Data Mining mit urheberrechtlich geschützten Textbeständen.“ *Zeitschrift für digitale Geisteswissenschaften (ZfdG)* 5 (2020). http://dx.doi.org/10.17175/2020_006.