ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Zeitschrift für digitale Geisteswissenschaften

Titel:
Vorzüge von Auszügen – Urheberrechtlich geschützte Texte in den digitalen Geisteswissenschaften (nach-)nutzen

Autor*in:
Melanie Andresen

Kontakt: melanie.andresen@ims.uni-stuttgart.de
Institution: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung
GND: 1143061535 ORCID: 0000-0002-3913-1273

Autor*in:
Markus Gärtner

Kontakt: markus.gaertner@ims.uni-stuttgart.de
Institution: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung
GND: 1268973939 ORCID: 0000-0002-2687-4350

Autor*in:
Sibylle Hermann

Kontakt: sibylle.hermann@ub.uni-stuttgart.de
Institution: Universitätsbibliothek Stuttgart
GND: 1073989070 ORCID: 0000-0001-9239-8789

Autor*in:
Janina Jacke

Kontakt: janina.jacke@uni-goettingen.de
Institution: Georg August Universität Göttingen, Seminar für Deutsche Philologie
GND: 108423968X ORCID: 0000-0001-7217-3136

Autor*in:
Nora Ketschik

Kontakt: nora.ketschik@ims.uni-stuttgart.de
Institution: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung
GND: 1268974390 ORCID: 0000-0001-8758-5432

Autor*in:
Felicitas Kleinkopf

Kontakt: felicitas.kleinkopf@kit.edu
Institution: Karlsruher Institut für Technologie, Institut für Informations- und Wirtschaftsrecht, Zentrum für Angewandte Rechtswissenschaft
GND: 1268974757 ORCID: 0000-0001-8670-2668

Autor*in:
Jonas Kuhn

Kontakt: jonas.kuhn@ims.uni-stuttgart.de
Institution: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung
GND: 1064354289 ORCID: 0000-0003-2860-5960

Autor*in:
Axel Pichler

Kontakt: axel.pichler@alumni.uni-graz.at
Institution: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung
GND: 14316130X ORCID: 0000-0002-9177-7645

DOI des Artikels:
10.17175/2022_007_v2

Nachweis im OPAC der Herzog August Bibliothek:
1845597966

Erstveröffentlichung:
03.11.2022

Version 2.0:
22.06.2023

Letzte Überprüfung aller Verweise:
06.09.2022

Format:
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:
Urheberrecht | Data Mining | Infrastruktur | Open Science | Digital Humanities | Forschungsdaten |

Empfohlene Zitierweise:
Melanie Andresen / Markus Gärtner / Sibylle Hermann / Janina Jacke / Nora Ketschik / Felicitas Kleinkopf / Jonas Kuhn / Axel Pichler: Vorzüge von Auszügen – Urheberrechtlich geschützte Texte in den digitalen Geisteswissenschaften (nach-)nutzen. In: Zeitschrift für digitale Geisteswissenschaften 7 (2022). 03.11.2022. Version 2.0 vom 22.06.2023. HTML / XML / PDF. DOI: 10.17175/2022_007_v2.

Version 2.0 (22.06.2023):
Korrekturen in Text und Fußnoten anhand der Monita in den Gutachten. Ergänzungen in der Bibliografie.

Melanie Andresen, Markus Gärtner, Sibylle Hermann, Janina Jacke, Nora Ketschik, Felicitas Kleinkopf, Jonas Kuhn, Axel Pichler

Vorzüge von Auszügen – Urheberrechtlich geschützte Texte in den digitalen Geisteswissenschaften (nach-)nutzen

Abstracts

Um urheberrechtlichen Einschränkungen beim Austausch zu Forschungsergebnissen von vornherein aus dem Weg zu gehen, konzentrieren sich viele digitale Geisteswissenschaftler*innen auf gemeinfreie Texte. Zur Überwindung dieser problematischen Beschneidung des Gegenstandsbereichs schlagen Schöch et al. 2020 sogenannte abgeleitete Textformate vor, die digitale Analyseverfahren unterstützen, den Text jedoch aus urheberrechtlicher Perspektive hinreichend verfremden. Das Projekt XSample entwickelt eine komplementäre Lösung, die die Berechtigung zur Weitergabe von Auszügen aus geschützten Texten (im Klartext) nutzt. Der forschungsgesteuerte Ansatz ermöglicht Gruppen, die an einer Nachnutzung interessiert sind, eine Optimierung des maximal erlaubten Auszugsvolumens entlang eigener Forschungsfragen.

In order to avoid copyright restrictions on the exchange of research results from the outset, many researchers in the digital humanities focus on texts in the public domain. To overcome this problematic limitation of the subject matter, Schöch et al. 2020 propose so-called derived text formats that support digital analysis procedures but sufficiently alienate the text from a copyright perspective. The XSample project is developing a complementary solution that leverages permission to share excerpts from copyrighted text (in plain text). The research-driven approach allows groups interested in reuse to optimize the maximum allowed excerpt volume along their own research questions.

1. Einleitung

Die korpusorientierte Forschung an Texten und anderen Materialien in den digitalen Geisteswissenschaften[1] ist durch das deutsche Urheberrecht eingeschränkt, das die Weitergabe von Forschungsdaten zu geschützten Werken und deren Archivierung nur in engen Grenzen erlaubt.[2] Die Restriktionen erschweren die Einhaltung der guten wissenschaftlichen Praxis sowie der *FAIR-Prinzipien* für Forschungsdateninfrastrukturen.[3] In vielen Projekten wird deshalb aus pragmatischen Gründen gänzlich darauf verzichtet, urheberrechtlich geschützte Texte einzubeziehen. Nicht selten bedeutet dies, dass zeitgenössische Texte – und mit ihnen bestimmte Fragestellungen – nahezu vollständig ausgeblendet werden (müssen).[4] Eine so weitreichende Beschneidung des Gegenstandsbereichs führt jedoch nicht nur zu Verzerrungen der Forschungslandschaft, die aus geisteswissenschaftlich-fachlichen Gründen problematisch sind, sie hat häufig auch zur Folge, dass sich die Entwicklung algorithmischer Verfahren mit sehr kleinen Datensätzen und / oder einer heterogenen Kombination von Quellen behelfen muss.

Selbstverständlich besteht für Forscher*innen die Möglichkeit, für die Forschung auf geschützten Texten in Verhandlungen mit den Rechteinhaber*innen zu treten. Optimal ist, wenn im Vorfeld eine Lizenzvereinbarung ausgehandelt werden kann, die eine unbegrenzte Weitergabe der geschützten Texte an Dritte zu Forschungszwecken einschließt; dies gelingt jedoch nicht in allen Fällen und erfordert einen erheblichen Aufwand und zeitlichen Vorlauf in Bezug auf jede zu verwendende Datenquelle. Damit ist das Vorgehen mit der Praxis datenintensiver Computermodellierung nur bedingt vereinbar, bei der etwa explorative Experimente zu unterschiedlichen Erweiterungen der Datenbasis nur bei einem Bruchteil der Daten eine längerfristige Weiterverfolgung bewirken.

[1] Die Namen der Autor*innen sind in den Verfasserangaben alphabetisch aufgelistet. Im Projekt befanden sich juristische Fragestellungen im Arbeitsschwerpunkt von Felicitas Kleinkopf; Markus Gärtner befasste sich mit der technischen Umsetzung der Infrastruktur; das erste Nutzungsszenario wurde von Melanie Andresen und Axel Pichler, das zweite von Janina Jacke und Nora Ketschik bearbeitet; Sibylle Hermann koordinierte die Projektarbeit und die Anbindung an die bibliothekarische Infrastruktur; Jonas Kuhn war für konzeptionelle Fragen verantwortlich. Die textuelle Darstellung in diesem Artikel wurde gemeinschaftlich von den Projektbeteiligten des XSample-Projekts, auch über die Zuständigkeitsgrenzen im Projekt hinweg, erstellt.

[2] Ähnliche Restriktionen gibt es auch in anderen Rechtsordnungen, das betrifft aufgrund derselben zugrunde liegenden EU-Urheberrechts-Richtlinien insbesondere die EU-Mitgliedsstaaten, während insbesondere das US-amerikanische Copyright-Law mit der sogenannten Fair-Use-Doktrin grundsätzlich anders ausgestaltet ist. Gegenstand dieser Darstellung ist allerdings allein das deutsche Urheberrecht inklusive seiner Grundlagen aus dem Unionsrecht.

[3] Die FAIR-Prinzipien formulieren vier zentrale Anforderungen an Forschungsdaten: Sie sollten Findable, Accessible, Interoperable und Reusable sein, siehe Wilkinson et al. 2016.

[4] Der Schutz von Texten durch das deutsche Urheberrecht endet siebzig Jahre nach Tod der Autor*innen, sodass kein direkter Zusammenhang zwischen Publikationsjahr und dem Ende des urheberrechtlichen Schutzes besteht.

Dieser Artikel sieht ein Desiderat für die digitalen Geisteswissenschaften – sei es bei der Erschließung eines Gegenstandsbereichs oder bei der Methodenentwicklung – darin, eine Forschungsdateninfrastruktur einzurichten, die ein exploratives Vorgehen unterstützt, sodass die Urheberrechtsfrage nicht länger per se ein Ausschlusskriterium für die Verwendung eines Textes oder eines Textkorpus ist. Neben dem langfristigen (politischen) Ziel einer verbesserten urheberrechtlichen Ausgangslage für die Forschung sollte dafür der bestehende rechtliche Korridor für eine Weitergabe von Forschungsergebnissen zu geschützten Texten ausgenutzt werden.

Schöch et al. schlagen zum Umgang mit der bestehenden Rechtslage eine Konvertierung der Texte in sogenannte abgeleitete Formate vor, welche für eine Reihe von digitalen Analyseverfahren geeignet sind, die den Text jedoch aus urheberrechtlicher Perspektive hinreichend verfremden.[5] Diese abgeleiteten Textformate halten beispielsweise für Textsegmente wie Kapitel oder Abschnitte lediglich die Häufigkeit der enthaltenen Einzelwörter oder *n-Gramme* (also kurzen Wortsequenzen) fest. Gängige Verfahren der Makroanalyse,[6] die etwa lexikalische Indikatoren für die Dynamik des Textverlaufs heranziehen, können auf dieser Basis zur Anwendung kommen. Der Urheberrechtsschutz wird dabei durch den Aufbruch der Textstruktur aufgehoben, sodass Restriktionen zur Archivierung, Weitergabe und Veröffentlichung der Datensätze nicht mehr zum Tragen kommen. Das Konzept der abgeleiteten Textformate leistet somit einen großen Beitrag zur Replizierbarkeit von Forschung und Nachnutzbarkeit von Forschungsdaten.

Allerdings stößt das Konzept der abgeleiteten Formate dort an seine Grenzen, wo die eigentliche Textgestalt forschungsrelevant wird. Dies ist nicht nur bei einer Mikroanalyse, also etwa beim *Close Reading*, der Fall (welches ohne Frage einen urheberrechtlich geklärten Gesamtzugriff auf den Text voraussetzt). Vor dem Hintergrund geisteswissenschaftlicher Fragestellungen macht häufig auch die Interpretation von Ergebnissen einer aggregierenden Makroanalyse den Zugriff auf einige relevante Textpassagen in ihrer Gesamtgestalt erforderlich.

Das XSample-Projekt hat daher einen Ansatz entwickelt, der komplementär zum Konzept der abgeleiteten Textformate eingesetzt werden kann (vgl. Abbildung 1). Dieser Ansatz nutzt das bestehende Recht zur Weitergabe von prozentual begrenzten Auszügen geschützter Werke zu Zwecken der wissenschaftlichen Forschung (§ 60c UrhG) und überträgt dies auf die Herausgabe von Korpusauszügen. Der erlaubte Umfang dieser Auszüge beträgt zwar in der Regel nur 15 Prozent eines Werks, der XSample-Ansatz ermöglicht jedoch eine dynamische, auf das individuelle Forschungsanliegen zugeschnittene Auswahl der ›hilfreichsten 15 Prozent‹. Hierzu können Nachnutzer*innen in Suchanfragen auf den Texten und gegebenenfalls vorhandenen Annotationen genau spezifizieren, welche Teile des Korpus für sie relevant sind. Dabei stellt der XSample-Ansatz sicher, dass die geschützten Primärdaten bei der Modellierung der Suchanfrage für die Nachnutzer*innen nicht einsehbar sind. Auf diese Weise werden die rechtlichen Möglichkeiten in einer zielführenden Art und Weise ausgeschöpft und nachhaltige Forschung mit urheberrechtlich geschützten Texten begünstigt, ohne den Urheberrechtsschutz aufzuheben. Insbesondere können auch Forschungsfragen bearbeitet werden, die den Rückgriff auf den exakten Wortlaut ausgewählter Textpassagen und umfangreichere Kontexte erfordern. Eine besondere Rolle nehmen dabei die Forschungsinfrastruktureinrichtungen ein, die Wissenschaftler*innen auf institutioneller Ebene unterstützen, indem sie ihnen digitale Werkzeuge, die z. B. der Verwaltung und Veröffentlichung ihrer Forschungsdaten dienen, zur Verfügung stellen. Zu diesen Forschungsinfrastruktureinrichtungen zählen heute vorwiegend die wissenschaftlichen Bibliotheken. Das im Projekt entwickelte Tool wird exemplarisch an die lokale Infrastruktur der Universität Stuttgart und das dort vorhandene Forschungsdatenrepositorium angebunden. Die im Projekt entwickelte Software steht frei zur Verfügung, sodass sie und die nötige Infrastruktur anderen Forschungsinfrastruktureinrichtungen bereitgestellt werden können.

[5] Vgl. Schöch et al. 2020.
[6] Vgl. Jockers 2013.

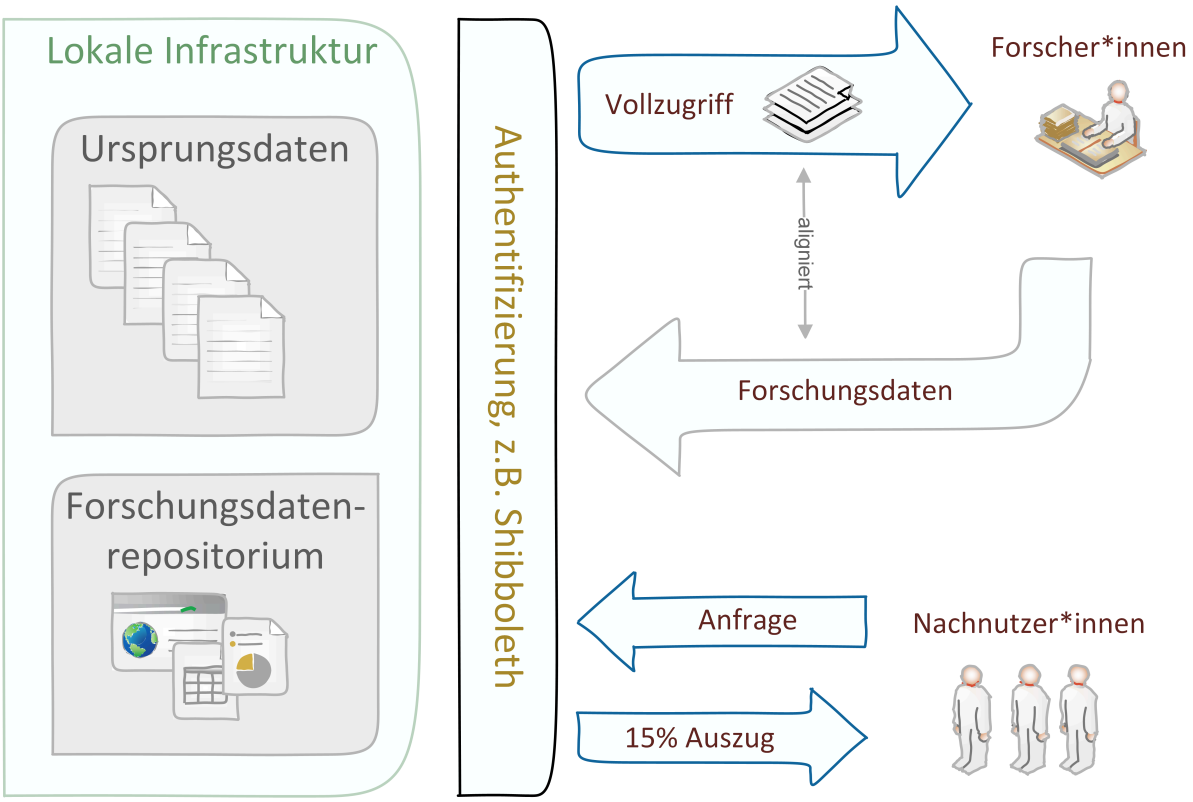

Abb. 1: In XSample entwickeltes Auszugskonzept. [Gärtner 2021]

Der vorliegende Beitrag geht in Kapitel 2 zunächst auf die rechtlichen Grundlagen ein, die für Verfahren des *Text- und Data-Mining* (TDM)[7] gelten und die Rahmenbedingungen für das hier präsentierte Auszugskonzept schaffen. Kapitel 3 stellt dar, wie die Verfügbarmachung von individuell zugeschnittenen Textauszügen innerhalb des rechtlichen Rahmens technisch umgesetzt werden kann. Dabei werden insbesondere die unterschiedlichen Akteur*innen in den Blick genommen sowie Schritte der Datenvorbereitung und verschiedene Möglichkeiten der Auszugsgenerierung vorgestellt. Anschließend werden in Kapitel 4 zwei Nutzungsszenarien präsentiert, die im Kontext konkreter Forschungsfragen ausführen, inwieweit abgeleitete Textformate und / oder das Auszugskonzeptkonzept für die Reproduktion ihrer Forschung und die Nachnutzung ihrer Daten praktikabel sind. Das erste Nutzungsszenario widmet sich der Wissenschaftssprache der geisteswissenschaftlichen Fächer Literaturwissenschaft, Linguistik und Philosophie und erstellt dazu ein Korpus aus insgesamt 135 urheberrechtlich geschützten Zeitschriftenartikeln. Das zweite Nutzungsszenario beschäftigt sich mit dem Phänomen des unzuverlässigen Erzählens, das in einem ersten Zugang anhand eines Korpus aus acht deutschsprachigen fiktionalen Erzählungen aus dem 19. bis zum 21. Jahrhundert untersucht wird, die teilweise dem Urheberrecht unterliegen. Nach Abschluss der Forschung sollen die Daten für die Überprüfung der Ergebnisse und zur Nachnutzung in weiteren Projekten zur Verfügung gestellt werden. Das Fazit in Kapitel 5 fasst die Ergebnisse zusammen und leitet praktische Handlungsempfehlungen und Desiderate ab.

2. Urheberrechtliche Rahmenbedingungen für das Text- und Data-Mining

Bei der Beforschung insbesondere neuerer Texte und Korpora müssen sich die digitalen Geisteswissenschaften mit Fragen des Urheberrechts auseinandersetzen. Der urheberrechtliche Rahmen für die Forschung mit TDM hat sich in den letzten Jahren mehrfach geändert, was es zusätzlich erschwert, aus geisteswissenschaftlicher Perspektive zu überblicken, wie bzw. in welchem Umfang mit geschützten Werken geforscht werden darf. Die letzten Änderungen ergaben sich im Juni 2021, als die *Richtlinie zum Urheberrecht im digitalen Binnenmarkt* (Digital Single Market-, kurz DSM-Richtlinie) im *Urheberrechtsgesetz* (UrhG) umgesetzt wurde.

[7] Unter TDM versteht das Urheberrecht »die automatisierte Analyse von einzelnen oder mehreren digitalen oder digitalisierten Werken, um daraus Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen« (§§ 44b Abs. 1, 60d Abs. 1 UrhG). Unterschieden wird rechtlich zwischen der eigentlichen automatisierten Analyse, die urheberrechtlich freigestellt ist, und den dafür notwendigen Vorbereitungsschritten, die wiederum urheberrechtlich relevant sind. Unter das gesetzgeberische Verständnis von TDM kann ein Großteil textbasierter Forschung gefasst werden, auch wenn die Forscher*innen ihre Analyse selbst möglicherweise nicht als Text- und Data-Mining bezeichnen würden.

Um einen kurzen Überblick über die Entwicklung der Gesetzeslage zu geben, wird nachfolgend (Kapitel 2.1) skizziert, in welchem Umfang Vervielfältigungen (§ 16 UrhG) und öffentliche Zugänglichmachungen (§ 19a UrhG) für die Erstellung und (gemeinsame) Beforschung von Korpora im Kontext von TDM in den verschiedenen Gesetzesfassungen erlaubt waren bzw. sind.[8] Daran anschließend werden Fragen nach der Zugänglichmachung und Nachnutzbarkeit von Korpora sowie mögliche Lösungsansätze diskutiert (Kapitel 2.2.), darunter insbesondere die rechtliche Grundlage für das in diesem Artikel vorgestellte Auszugskonzept. Abschließend werden die wesentlichen Anforderungen an Forschungsinfrastruktureinrichtungen zusammengefasst, die sich aus den juristischen Rahmenbedingungen ergeben (Kapitel 2.3).

2.1. Die Gesetzesentwicklung

Bis zum 1. März 2018 enthielt das UrhG keine gesonderte Erlaubnis, Werke zu Zwecken des TDM zu nutzen. Deswegen waren TDM-Analysen an urheberrechtlich geschützten Werken nur insoweit möglich, wie die Werke nicht kopiert oder weitergegeben wurden: Diese Handlungen sind nach den Regelungen in §§ 16, 19a UrhG urheberrechtlich relevant, die das ausschließliche Recht des Urhebers enthalten, sein Werk zu vervielfältigen und es öffentlich zugänglich zu machen. Eine ausdrückliche Erlaubnis[9] dieser Handlungen wurde im Jahr 2018 durch das Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (UrhWissG) mit einem neuen § 60d UrhG geschaffen. Unter der Voraussetzung des rechtmäßigen Zugriffs erlaubt § 60d UrhG der nicht-kommerziellen wissenschaftlichen Forschung seither die Vervielfältigung (§ 16 Abs. 1 UrhG) und die öffentliche Zugänglichmachung (§ 19a UrhG) geschützter Werke zu Zwecken des TDM. Seitdem existiert im deutschen Urheberrecht erstmalig eine explizite Erlaubnis, geschützte Werke für das TDM auch umfangreich zu verarbeiten. Defizitär war bereits zu jenem Zeitpunkt, dass die Korpora zwar bei bestimmten Einrichtungen, darunter auch den in diesem Artikel adressierten Forschungsinfrastruktureinrichtungen, aufbewahrt werden durften, aber keine Möglichkeit bestand, die Korpora zu Zwecken von Anschlussforschungen nachzunutzen.

Aufgrund des *Gesetzes zum Urheberrecht im digitalen Binnenmarkt* vom 7. Juni 2021[10] hat sich der Rechtsrahmen ein weiteres Mal verändert. Seitdem darf Forschung mit TDM auch zu anderen Zwecken als zur nicht-kommerziellen wissenschaftlichen Forschung (diese ist weiterhin in § 60d UrhG geregelt), z. B. in Unternehmen oder in der Presse, praktiziert werden, wenn auch in eingeschränktem Umfang (§ 44b UrhG).[11] Die Erlaubnis ist dabei nach wie vor an den rechtmäßigen Zugang gebunden. Dieser rechtmäßige Zugang liegt dann vor, wenn die genutzten Texte in Buchform vorliegen, im Internet frei abrufbar sind oder als E-Books lizenziert wurden.[12] Neu ist auch, dass jedenfalls für Vervielfältigungen im Rahmen von TDM nunmehr keine Vergütung mehr anfällt (§ 60h Abs. 2 Nr. 3 UrhG). Diese war zuvor durch die jeweiligen Einrichtungen zu tragen.

2.2. Die Erlaubnisse für Forschungsinfrastruktureinrichtungen

§ 60d UrhG in seiner aktuellen Fassung erlaubt t es bestimmten privilegierten Personenkreisen und Einrichtungen, vollständige Werke zu Zwecken des TDM zu vervielfältigen (Abs. 1 bis 3) und die Vervielfältigungen (d. h. nicht die unveränderten Ursprungsmaterialien) für die gemeinsame wissenschaftliche Forschung bestimmt abgegrenzten Personenkreisen sowie einzelnen Dritten zur Überprüfung der wissenschaftlichen Forschung öffentlich zugänglich zu machen (Abs. 4 S. 1). Im Gegensatz zu seiner Vorgängernorm knüpft § 60d UrhG in seiner neuen Fassung nicht allgemein an nicht-kommerzielle wissenschaftliche Zwecke[13] an, sondern berechtigt Forschungsorganisationen (Hochschulen, Forschungsinstitute und sonstige Einrichtungen, die wissenschaftliche Forschung betreiben, hierzu zählen auch die in diesem Beitrag adressierten Forschungsinfrastruktureinrichtungen), sofern diese 1. nicht kommerzielle Zwecke verfolgen, 2. sämtliche Gewinne in die Forschung reinvestieren oder 3. im Rahmen eines staatlich anerkannten Auftrags im öffentlichen Interesse tätig sind. Public-Private-Partnerships, d. h. Kooperationen mit privaten Unternehmen, sind nur dann erfasst, wenn letztere keinen bestimmenden Einfluss auf die Forschungsorganisation haben und keinen bevorzugten Zugang zu den Forschungsergebnissen erhalten (§ 60d Abs. 2 S. 3 UrhG). § 60d Abs. 3 Nr. 1 UrhG benennt nunmehr auch ausdrücklich sogenannte Kulturerbe-

[8] Einen umfassenderen Überblick über die Detailfragen des Forschungsprozesses bieten zum alten Recht Dreier / Schulze 2018, § 60d; Kleinkopf et al. 2021; vorwiegend auch Schöch et al. 2020, Absatz 5–14; zu § 60d in seiner neuen Fassung, vgl. Dreier in Dreier / Schulze 2022, § 44b und § 60d; Raue 2021; Kleinkopf / Pflüger 2021, S. 645–647; eine Betrachtung auf EU-Ebene bieten Gärtner et al. 2021, S. 11–13. Vgl. auch Kleinkopf 2022.
[9] Erlaubnisnormen werden im deutschen Urheberrecht als »Schranken« bzw. »Schrankenbestimmungen« bezeichnet.
[10] Bundesgesetzblatt Jahrgang 2021 Teil I Nr. 27, ausgegeben zu Bonn am 4. Juni 2021.
[11] In diesen kommerziellen Kontexten dürfen für das TDM Werke vervielfältigt werden (§ 44b Abs. 2 S. 1 UrhG), diese müssen aber gelöscht werden, wenn sie für das TDM nicht mehr erforderlich sind (§ 44b Abs. 2 S. 2 UrhG). Zudem können Rechteinhaber*innen an ihren Werken (maschinenlesbare) Nutzungsvorbehalte anbringen, die von denjenigen, die auf Grundlage des § 44b UrhG TDM betreiben, ab dem Zeitpunkt ihrer Erklärung beachtet werden müssen (§ 44b Abs. 3 S. 1, 2 UrhG). Im Rahmen des § 44b UrhG ist nicht erlaubt, Werke oder Werkteile öffentlich zugänglich zu machen, auch nicht an bestimmt abgegrenzte Personenkreise. Aufgrund der Löschpflicht dürfen die erstellten Korpora auch nicht längerfristig aufbewahrt werden. Für wissenschaftliche Zwecke ist es deswegen unbedingt notwendig, sich auf § 60d UrhG berufen zu können.
[12] Vertraglich oder technisch darf das TDM im Rahmen wissenschaftlicher Zwecke (§ 60d UrhG) auch im Grundsatz nicht ausgeschlossen werden (§§ 60g Abs. 1, 95b Abs. 1, 3 UrhG). Etwas anderes gilt für kommerzielles TDM im Rahmen des § 44b UrhG oder für die Sicherung der Funktionsfähigkeit technischer Systeme (§ 60d Abs. 6 UrhG).
[13] Vgl. zum genauen Verständnis der nicht-kommerziellen Zwecke die Ausführungen in Absatz 19.

Einrichtungen wie Bibliotheken, Museen, Archive und Einrichtungen im Bereich des Ton- und Filmerbes, wobei hier nicht an nicht-kommerzielle Zwecke, sondern an ihre öffentliche Zugänglichkeit angeknüpft wird. Individualforscher*innen sind weiterhin unter der Voraussetzung, dass sie nicht-kommerzielle Zwecke verfolgen, erfasst (§ 60d Abs. 3 Nr. 2 UrhG). Die öffentliche Zugänglichmachung steht unter der Voraussetzung der nicht-kommerziellen Zwecke, sodass auch die öffentlich zugänglichen Kulturerbe-Einrichtungen diese letztlich beachten müssen, um vom vollen Umfang der Erlaubnisse des § 60d UrhG zu profitieren. Die Zugänglichmachung ist jeweils zu beenden, wenn die gemeinsame Forschung oder die Überprüfung abgeschlossen ist (Abs. 4 S. 2).

Die Befugnis zur Weitergabe der Korpora während der Forschungsarbeiten ist also deutlich enger gefasst, als es ein allgemeiner Verweis auf die öffentliche Zugänglichmachung (§ 19a UrhG) zur Folge hätte: Sie erlaubt die öffentliche Zugänglichmachung eines Werks in einer Weise, dass es »Mitgliedern der Öffentlichkeit von Orten und zu Zeiten ihrer Wahl zugänglich ist«. Das bedeutet, dass das Werk Personen zugänglich gemacht wird, die der Öffentlichkeit angehören, d. h. zu denen keine persönliche Beziehung existiert, vgl. § 15 Abs. 3 UrhG, und die einer »unbestimmten Zahl potentieller Adressaten« und einer »ziemlich großen Zahl von Personen« angehören.[14] Die Erlaubnis, Korpora während der TDM-Forschungsarbeiten weiterzugeben (§ 60d Abs. 4 S. 1 UrhG), bezieht sich hingegen nur auf bestimmt abgegrenzte, d. h. weiter eingegrenzte Personenkreise, die der Öffentlichkeit angehören. Bei Forschungsgruppen handelt es sich in der Regel nicht um eine Öffentlichkeit,[15] weswegen es auf diese Erlaubnis in den meisten Fällen nicht ankommt.

Die TDM-Korpora dürfen gemäß § 60d UrhG so lange aufbewahrt werden, wie es für die Forschung oder für Überprüfungszwecke erforderlich ist. Die Dauer erfasst im Regelfall die von der guten wissenschaftlichen Praxis geforderten zehn Jahre,[16] sie kann im Einzelfall aber auch kürzer oder länger bemessen sein. Aufgrund der Wissenschaftsfreiheit unterliegt die Einschätzung der Aufbewahrungsdauer den Forschenden und ist nur eingeschränkt überprüfbar.[17]

Eine Erlaubnis, die Korpora zur Aufbewahrung an Forschungsinfrastruktureinrichtungen weiterzugeben, ist nicht mehr ausdrücklich enthalten. Die Weitergabemöglichkeit kann allenfalls mittels einer extensiven Auslegung aus der Gesetzesbegründung abgeleitet werden, denn diese scheint von einer Aufbewahrung durch Kulturerbe-Einrichtungen auszugehen:[18] »Hiernach kann auch eine dauerhafte Speicherung erforderlich und folglich zulässig sein, insbesondere, wenn sie durch Kulturerbe-Einrichtungen und nicht durch die Forschungseinrichtung selbst erfolgt«.[19] Möglich ist aber in jedem Fall, die TDM-Projekte gleich von Beginn an auf zentralen Bibliotheksservern bzw. einrichtungsübergreifenden Servern anzusiedeln, sodass dort die Archivierung ohne eine gesonderte Übermittlung erfolgen kann. Dann bewegt man sich im Rahmen dessen, was die Gesetzesbegründung ausdrücklich enthält, und vermeidet gleichzeitig, das Korpus erneut zu vervielfältigen (§ 16 UrhG).

Unklar ist aber weiterhin, ob zu Zwecken von Anschlussforschungen auf die Korpora zugegriffen werden kann. Gemäß § 60d Abs. 5 UrhG dürfen die Korpora (nach dem Gesetz jedoch nicht die unveränderten Ursprungsdaten)[20] so lange aufbewahrt werden, wie es für Zwecke der Überprüfung der Qualität der wissenschaftlichen Forschung oder für die Forschung selbst erforderlich ist. Das impliziert, dass es auch abseits der Überprüfung wissenschaftliches Interesse an den Korpora geben kann. Auch die DSM-Richtlinie setzt voraus, dass nach Abschluss der Forschungsarbeiten bzw. während der Langzeitarchivierung noch Interesse an weiterer Beforschung der Korpora bestehen kann:[21] »Die Nutzung zum Zwecke der wissenschaftlichen Forschung außerhalb des Text und Data Mining, etwa die Begutachtung unter wissenschaftlichen Fachkollegen und gemeinsame Forschungsarbeiten, sollte nach wie vor unter die Ausnahme oder Beschränkung im Sinne von Artikel 5 Absatz 3 Buchstabe a der Richtlinie 2001/29/EG fallen, sofern diese Bestimmung anwendbar ist.« [22] Für eine Nachnutzung der Korpora kommen neben den von Schöch et al. vorgestellten abgeleiteten Textformaten sogenannte *Closed-Room-Zugänge* [23] (§§ 60e Abs. 4, 60f UrhG) und die Erlaubnis der auszugsweisen Nutzung zu Zwecken der wissenschaftlichen Forschung (§ 60c UrhG) in Betracht. Nachfolgend werden diese beiden Optionen dargestellt und gegeneinander abgewogen.

[14] Dreier in Dreier / Schulze 2022, § 15 Randnummer 38 mit Verweis auf den EuGH.
[15] Vgl. Raue 2021, S. 799.
[16] Vgl. Leitlinien zur Sicherung der guten wissenschaftlichen Praxis, Deutsche Forschungsgemeinschaft 2019, Leitlinie 17.
[17] Vgl. Raue 2021, S. 799.
[18] Vgl. Kleinkopf / Pflüger 2021, S. 647.
[19] Bundestagsdrucksache 19/27426, S. 97.
[20] Etwas anderes kann gelten, wenn die Ursprungsdaten entsprechend lizenziert sind, vgl. dazu auch Kapitel 3.2.
[21] Vgl. Kleinkopf / Pflüger 2021, S. 647.
[22] Erwägungsgrund 15 S. 5 DSM-Richtlinie.
[23] Vgl. Schöch et al. 2020, Absatz 4f.

§§ 60e Abs. 4, 60f UrhG erlauben es Bibliotheken und anderen Kulturerbe-Einrichtungen wie öffentlich zugänglichen Museen, ihren Nutzer*innen Werke aus ihrem Bestand an Terminals in ihren Räumen für deren Forschung oder private Studien zugänglich zu machen (sogenannte *Terminal-Schranke*). Die Nutzer*innen dürfen sogenannte *Anschlusskopien* [24] im Umfang von zehn Prozent erstellen. Einzelne Werke geringen Umfangs[25] wie Beiträge aus wissenschaftlichen Zeitschriften dürfen hingegen vollständig genutzt werden. Auf Grundlage des § 60e Abs. 4 UrhG kann also durch Kulturerbe-Einrichtungen Vollzugriff gewährt werden, außerdem können die Anschlusskopien interessengerecht erstellt werden. Ein entscheidender Nachteil an § 60e Abs. 4 UrhG ist gleichwohl, dass jedenfalls der erste Zugriff auf die Werke nur an Terminals vor Ort erfolgen kann (sogenannte Closed-Room-Zugänge).[26]

Anders verhält es sich mit § 60c UrhG, der Erlaubnisnorm für Zwecke der nicht-kommerziellen wissenschaftlichen Forschung, auf den sich das hier vorgestellte Konzept stützt. § 60c UrhG basiert auf Art. 5 Abs. 3 lit. a InfoSoc-Richtlinie, der den EU-Mitgliedsstaaten ermöglicht, in ihrem nationalen Urheberrecht Erlaubnisse von Vervielfältigungen und öffentlicher Zugänglichmachung »für Zwecke der wissenschaftlichen Forschung, sofern – außer in Fällen, in denen sich das als unmöglich erweist – die Quelle, einschließlich des Namens des Urhebers, wann immer das möglich ist, angegeben wird und soweit das zur Verfolgung nicht kommerzieller Zwecke gerechtfertigt ist« vorzusehen.

§ 60c Abs. 1 Nr. 1 UrhG erlaubt es, zu Zwecken der nicht-kommerziellen wissenschaftlichen Forschung bis zu 15 Prozent von Werken und auch vollständige Werke geringen Umfangs zu vervielfältigen und an bestimmt abgegrenzte Personenkreise für deren eigene wissenschaftliche Forschung öffentlich zugänglich zu machen, d. h. weiterzugeben. Auf dieser Grundlage können geschützte Werke auf individuelle Anfrage teils vollständig, teils auszugsweise, weitergegeben werden, auch digital und ohne Ortsbindung. Nicht erlaubt ist allerdings, Werke für eine gesamte Einrichtung frei abrufbar zu machen.[27] Die nicht-kommerziellen Zwecke können auch bei Drittmittelforschung sowie dann vorliegen, wenn Forschende ihre Ergebnisse in einem Verlag veröffentlichen und Honorare erhalten, kommerzielle Zwecke sind aber jedenfalls dann anzunehmen, wenn Forschung betrieben wird, um Waren oder Dienstleistungen zu entwickeln und diese zu vermarkten.[28] Entscheidend ist bei der Bestimmung der nicht-kommerziellen Zwecke nicht die organisatorische Einrichtung oder Finanzierung, sondern, ob die jeweilige Nutzung auf Gewinnerzielung ausgerichtet ist.[29]

Wenn Forschende also nach § 60c UrhG Texte auszugsweise zur Nachnutzung erhalten und diese anschließend für TDM im Sinne des § 60d UrhG nutzen, werden zwei Erlaubnisnormen, die auf unterschiedlichen Richtlinien beruhen (nämlich einerseits der InfoSoc-Richtlinie und andererseits der DSM-Richtlinie), miteinander kombiniert. Das ist rechtlich möglich.[30]

Daneben entspricht die Nachnutzbarkeit auf Grundlage des § 60c UrhG auch den Interessen der Urheber*innen, das gilt insbesondere deswegen, weil § 60c UrhG vergütungspflichtig ist (§ 60h UrhG) und dadurch ein finanzieller Ausgleich für die Rechteinhaber*innen hergestellt wird, schließlich ist für eine Nutzung gemäß § 60d UrhG ein rechtmäßiger Zugang erforderlich, der zumeist einen finanziellen Ausgleich für den*die Urheber*in enthält. § 60c UrhG setzt diesen rechtmäßigen Zugang allerdings nicht voraus. Die Vergütung stellt also einen Ausgleich der urheberrechtlichen Interessen her. Insgesamt ermöglicht § 60c UrhG insofern eine flexiblere Korpus-Nachnutzung als §§ 60e Abs. 4, 60f UrhG, weswegen sich das in XSample entwickelte Konzept auf § 60c UrhG stützt.

2.3. Rahmenbedingungen für das Auszugskonzept

Zusammenfassend orientiert sich das hier vorgestellte Auszugskonzept an folgenden rechtlichen Einschränkungen, die von Forschungsinfrastruktureinrichtungen, die Korpora mit geschützten Texten bereitstellen möchten, zu prüfen sind:

Einbeziehung der Forschungsinfrastruktureinrichtungen: Forschungsinfrastruktureinrichtungen, die die Korpora aufbewahren und gegebenenfallsbereitstellen, sollten von Beginn an in die Projektkonzeptionen integriert werden.

[24] Der Terminus ›Anschlusskopie‹ meint, dass sich die Erlaubnis der Erstellung einer zehnprozentigen Kopie von Nutzer*innen an die Erlaubnis der Einrichtung, Werke an Terminals zugänglich zu machen, anschließt; vgl. dazu auch Dreier in Dreier / Schulze 2022, § 60e Randnummer 21-24.
[25] Unter ›geringem Umfang‹ werden gemeinhin 25 Seiten verstanden, vgl. Dreier in Dreier / Schulze 2022, § 60c Randnummer 15, § 60a Randnummer 22; Bundestagsdrucksache 18/12329, S. 35.
[26] Vgl. Schöch et al. 2020, Absatz 5.
[27] Das entspräche einer Einstellung in ein Universitäts-Intranet, diese ist jedoch gerade nicht erlaubt, vgl. Dreier in Dreier / Schulze 2022, § 60c Randnummer 9; Bundestagsdrucksache 15 / 837, S. 34.
[28] Vgl. Dreier in Dreier / Schulze 2022, § 60c Randnummer 6 mit Verweis auf die Gesetzesbegründung, Bundestagsdrucksache 18/12329, S. 39.
[29] Vgl. Dreier in Dreier / Schulze 2022, § 60a Randnummer 7.
[30] Nach der DSM-Richtlinie ist es zulässig, weitere Nutzungen der TDM-Korpora auf die ältere InfoSoc-Richtlinie zu stützen, das besagt zum einen Erwägungsgrund 15 S. 5 und zum anderen Art. 24 Abs. 2, 25 DSM-Richtlinie. Auch die Rechtsprechung hat bereits in der Vergangenheit urheberrechtliche Erlaubnisnormen miteinander kombiniert, wenn ihre jeweiligen Voraussetzungen erfüllt sind, EuGH GRUR 2014, 1078 – TU Darmstadt / Ulmer; BGH GRUR 2015, 1101 – Elektronische Leseplätze II; so bereits Kleinkopf et al. 2021, S. 198f.

Aufbewahrungsdauer: Forscher*innen, die die Korpora bereitstellen möchten, sollten eine der Forschung angemessenen Aufbewahrungsdauer für die Korpora vorschlagen. Wenn diese von den von der DFG vorgeschlagenen zehn Jahren[31] abweicht, sollte eine explizite Begründung erfolgen.

Inhalt der Korpora: Die gespeicherten und gegebenenfalls bereitgestellten Korpora dürfen nicht die unveränderten Ursprungsdaten enthalten, es sei denn, die betreffende Einrichtung hat hierzu entsprechende Lizenzen erworben.

Nachnutzung: Nachnutzer*innen müssen nicht-kommerzielle, wissenschaftliche Zwecke verfolgen. Die Verifizierung kann dadurch erfolgen, dass Interessierte sich in Bezug auf die Zugehörigkeit zu einer Forschungseinrichtung verifizieren und zudem versichern, die Daten nur für die nicht-kommerzielle wissenschaftliche Forschung zu verwenden.[32] Die Korpusauszüge dürfen nur bestimmt abgegrenzten Personenkreisen zugänglich gemacht werden, d. h. auf individuelle Anfrage.

Umfang der Korpusauszüge: Korpusauszüge dürfen maximal 15 Prozent ganzer Werke betragen, kurze Werke wie z. B. Aufsätze aus Zeitschriften (maximal 25 Seiten) können dagegen vollständig herausgegeben werden.

3. Technische Umsetzung

Wie im vorigen Kapitel beschrieben, bedient sich der in XSample verfolgte Ansatz der rechtlichen Erlaubnis, Auszüge bis zu einem Umfang von 15 Prozent eines Werks für Forschungszwecke weiterzugeben. Dieses Auszugskonzept wurde im Rahmen des Projekts prototypisch implementiert und zielt insbesondere darauf ab, den Nutzer*innen die für ihre Forschungsfrage ›hilfreichsten‹ Auszüge zu liefern. Die technische Umsetzung wird im Folgenden umrissen. Hierfür werden zunächst die Workflow-Akteur*innen (Kapitel 3.1) und die im Workflow eingesetzten Serverkomponenten (Kapitel 3.2) vorgestellt. Im Anschluss werden die Vorbereitungsschritte für die Erstellung eines Auszugs (Kapitel 3.3) sowie verschiedene Möglichkeiten der Auszugsgenerierung (Kapitel 3.4) erläutert. Für letztere liegt der Fokus auf der Verwendung im Korpus enthaltener Annotationen, um mittels Suchanfragen Auszüge zu erhalten, die optimal auf die Bedürfnisse der Nachnutzer*innen zugeschnitten sind. Abschließend wird die Nachhaltigkeit und Nachnutzbarkeit der hier vorgestellten Infrastruktur thematisiert (Kapitel 3.5).

3.1 Workflow-Akteur*innen

XSample unterscheidet zwischen drei Akteur*innen im Workflow: Erstens den Infrastrukturbetreiber*innen, zweitens den Datenanbieter*innen / -lieferant*innen und drittens den Nachnutzer*innen. Infrastrukturbetreiber*innen gehören zu den oben adressierten Forschungsinfrastruktureinrichtungen und bieten die infrastrukturelle Komponente zur Ablage und Verwaltung der verschiedenen im Workflow anfallenden Daten. Eine zentrale Bedeutung kommt hierbei der Authentifizierung von Nutzer*innen und einem feingranularen Rechtemanagement zu, um die beschriebenen rechtlichen Bedingungen (z. B. Zugriffsbeschränkung für bestimmt abgegrenzte Personenkreise) erfüllen zu können. Neben der reinen Datenablage dient das zugrundeliegende Repositorium mit seiner Weboberfläche gleichzeitig als zentraler Einstiegspunkt für die weiteren beteiligten Akteur*innen. Als Datenlieferant*innen werden im XSample-Kontext sämtliche Personen oder Personenkreise bezeichnet, welche (gemäß § 60d UrhG) TDM auf geschützten Werken zum Zwecke nicht-kommerzieller Forschung durchführen und die dabei erzeugten Korpora zur Nachnutzung durch XSample verfügbar machen möchten. Die Gruppe der Nachnutzer*innen schließlich beinhaltet die nach § 60c UrhG bestimmt abgegrenzten Personenkreise, denen auszugsweise Zugang zu geschützten Werken zum Zwecke nicht-kommerzieller Forschung gewährt werden darf.

3.2 Infrastruktur: Repositorium und Auszugsgenerierung

Innerhalb des XSample-Workflows kommen zwei getrennte Serverkomponenten zum Einsatz: das Repositorium und die Auszugsgenerierung. Beide stehen unter der Verwaltung der Infrastrukturbetreiber*innen und verfügen über jeweils eigene Weboberflächen.

[31] Leitlinien zur Sicherung der guten wissenschaftlichen Praxis, Deutsche Forschungsgemeinschaft 2019, Leitlinie 17.
[32] Den Einrichtungen werden auch beim Kopienversand keine weitergehenden Prüfpflichten auferlegt, es sei denn, es handelt sich um offensichtliche Missbrauchsfälle, vgl. Dreier in Dreier / Schulze 2022, § 60e Randnummer 17, 27, 28; Stieper in Schricker / Loewenheim 2020, § 60e Randnummer 37.

Das Repositorium dient vorwiegend der Ablage und Verwaltung der Korpus- und eventuell Ursprungsdaten und wird innerhalb des Prototyps durch eine Dataverse-Instanz realisiert. Die Dataverse Software ist ein Open-Source-Projekt auf Basis des JSF-Frameworks mit einer aktiven Community aus Entwickler*innen und regelmäßigen Nutzer*innen. Dataverse bietet die Möglichkeit, abgelegte Daten auf verschiedene Weise zu organisieren oder zu gruppieren, und verfügt überdies über eine Rechteverwaltung, die es erlaubt, bis auf die Ebene einzelner Datensätze zu entscheiden, ob ein komplett öffentlicher Zugang (*Public Domain*), das Teilen mit einzelnen Individuen oder Gruppen (*Shared Domain*), oder eine für andere uneinsehbare Ablage (*Private Domain*) gewünscht ist. Dies ist im Kontext von XSample besonders relevant, da zwar während eines laufenden Forschungsprojekts § 60d UrhG das Teilen der Daten innerhalb von Forschungsgruppen und zu Überprüfungszwecken erlaubt (Shared Domain), allerdings nach Projektende diese in ihrer Gesamtheit nicht mehr (ausdrücklich) öffentlich zugänglich gemacht oder geteilt werden dürfen (Private Domain). Zusätzlich unterscheidet Dataverse zwischen der Auffindbarkeit von Ressourcen und dem direkten Zugriff auf dieselben. Somit lassen sich die Metadaten als eigenständiger Datensatz veröffentlichen, während die eigentlichen Dateien innerhalb des Datensatzes aber vor jeglichem direkten Zugriff abgeschirmt sind.

Neben dem Repositorium stellt die Auszugsgenerierung als eigener Server die zweite Komponente im XSample-Workflow dar. Sie leitet Nutzer*innen auf einer Weboberfläche durch die individuelle Auszugskonfiguration und stellt am Ende des XSample-Workflows die im Auszug enthaltenen Daten als Download zur Verfügung. Dieser Server basiert ebenfalls auf JSF und kommuniziert mit der Dataverse-Instanz über eine Webschnittstelle, um auf dort abgelegte Ressourcen und Metadaten zuzugreifen. Um auf nicht-öffentliche Datensätze zugreifen zu können, benötigt der XSample-Server einen eigenen Account für das entsprechende Dataverse, der allerdings nur Leserechte beinhalten muss, denn der XSample-Server selbst schreibt oder modifiziert keine Daten im Repositorium. Zur Integration des Servers in die bestehende Dataverse-Infrastruktur wird dieser in der Dataverse-Instanz als sogenanntes external tool registriert. Diese Schnittstelle in Dataverse ermöglicht es, für bestimmte Dateitypen oder Datensätze externe Server zu registrieren, die den Nutzer*innen dann als zusätzliche Optionen neben Download oder Betrachtung angezeigt werden (vgl. Abbildung 2). Dies ermöglicht eine Integration der XSample-Komponenten, ohne Code-Modifikationen an Dataverse vornehmen zu müssen, und erlaubt überdies, komplett auf eine eigene Authentifizierung von Nutzer*innen von Seiten des XSample-Servers zu verzichten, da diese bereits bei Dataverse vorgeschaltet ist. Auf diese Weise werden die zur Nachverfolgung der Nutzer*innen nötigen Daten bei der Weiterleitung zum XSample-Server sogleich mit übermittelt.

3.3 Datenaufbereitung für die Auszugsgenerierung

Bevor Auszüge aus einem Korpus generiert werden können, sind mehrere Vorbereitungsschritte notwendig:

Zulässige Formate: Alle für die Auszüge zu verwendenden Dateien des Korpus müssen in einem nicht öffentlich zugreifbaren Bereich (Private Domain) des Repositoriums abgelegt werden. Im Kontext der Prototypenimplementierung ist bisher nur eine begrenzte Anzahl von Formaten für Ursprungsdaten (PDF, EPUB oder TXT) und Annotationen (TEI-Subset und *CoNLL*-ähnliche tabellarische Formate) vorgesehen. Während der Konzeptphase lag hierbei der Fokus auf EPUB und TXT Primärdaten, sowie Annotationen in einem Subset des weit verbreiteten TEI-Formats aus dem zweiten Nutzungsszenario (Kapitel 4.2). Bedingt durch zeitliche Überschneidungen verschob sich dieser Fokus im Verlauf der Implementierungsphase auf Daten des ersten Nutzungsszenarios (Kapitel 4.1), konkret auf PDF-Dateien und das tabellarische Format des CoNLL-2009-Shared-Tasks. Werden unveränderte Ursprungsdaten eingegliedert, ist von Seiten der Infrastrukturbetreiber*innen und Datenanbieter*innen auch zu prüfen, ob für die verwendeten Werke Archivierungsrechte oder Lizenzen mit äquivalenten Berechtigungen vorliegen, wie in Kapitel 2.3 beschrieben.

Alignierung: Neben den reinen Annotationen im Korpus muss auch eine Abbildung einzelner Annotationen auf die zugrunde liegenden Segmente der Ursprungsdaten geliefert werden, um beide innerhalb von XSample alignieren zu können. Da beispielsweise im Fall von Dateien im PDF-Format Auszüge zwangsläufig als Sammlung vollständiger Seiten erzeugt werden und die Zusammensetzung der Auszüge mittels Suchanfragen auf Basis der Annotationen gesteuert werden kann, müssen diese Annotationen (bzw. die dazugehörigen Suchergebnisse) auf die ursprünglichen Seiten abbildbar sein. Aktuell sieht der XSample-Prototyp für diese Alignierung entweder die Verwendung zusätzlicher Annotationsebenen direkt im Korpus vor, oder aber das Erstellen und Mitliefern zusätzlicher tabellarischer Dateien, welche eine simple Abbildung relevanter Segmentierungseinheiten seitens der Annotationen (z. B. Sätze) auf die primären Segmente der Ursprungsdaten enthält. Abhängig vom jeweiligen Projektinhalt und dem Anteil manueller Vorverarbeitung kann dieser Vorbereitungsschritt eine große Hürde darstellen. Dies sollte schon frühzeitig im Projekt eingeplant werden, damit die notwendigen und nicht selten ›format-fremden‹[33] Informationen

[33] Metainformationen wie Seitenzahlen auf der Ebene von Sätzen oder einzelnen Worten im Text sind in etablierten Annotationsschemata oder Formaten in der Regel nicht vorgesehen und erfordern somit zusätzlichen Aufwand, bzw. spezielle Anpassungen.

nicht im Forschungsprozess verloren gehen und entweder durch manuellen Zusatzaufwand oder Anpassung automatischer Verarbeitungsschritte wiederhergestellt werden müssen. Die Ablage der Alignierungsinformationen erfolgt analog zu obigen Ursprungs- und Annotationsdateien in einem nicht-öffentlichen Bereich.

Erzeugung eines Manifests: Als letztes muss ein sogenanntes XSample-Manifest (siehe Beispielcode unten) erzeugt und im Repositorium abgelegt werden. Dieses Manifest ist eine Datei im JSON-LD Format, die Metadaten zu den einzelnen für die weitere Verarbeitung relevanten Ressourcen im Korpus enthält. Primär bestehen diese Informationen aus Angaben zum Ablageort, Format und Umfang einzelner Dateien. Daneben sind aber auch Informationen zu Rechteinhaber*innen der einzelnen Werke im Korpus enthalten, damit der XSample-Server bei der Auszugsgenerierung der Namensnennungspflicht nachkommen kann. Im Manifest kann zusätzlich ein fixer Bereich für die statische Auszugsgenerierung festgelegt werden. Dies erlaubt es Datenlieferant*innen beispielsweise, besonders interessante Passagen als Teil des Standard-Auszugs zu definieren. Der Umfang des statisch definierten Auszugs muss nicht zwangsläufig die vollen 15 Prozent ausschöpfen, wenn interessante Inhalte auch in Auszügen geringeren Umfangs präsentiert werden können. Somit können Nachnutzer*innen den Rest ihrer Quote beispielsweise für zielgerichtete Varianten der Auszugserstellung nutzen. Derzeit erfolgt die Erstellung eines XSample-Manifests komplett händisch. Als zukünftige Erweiterung ist ein Assistent geplant, der Datenlieferant*innen auf der XSample-Webseite dabei helfen soll, Inhalte für Manifeste zu definieren, ohne direkt im JSON-LD-Format schreiben zu müssen. Im Gegensatz zu den restlichen Dateien muss das Manifest öffentlich auffindbar sein (Shared oder Public Domain), um den Nachnutzer*innen als Einstiegspunkt für die Auszugsgenerierung zu dienen. Da Dataverse für veröffentlichte Ressourcen DOIs (*Digital Object Identifier*) vergibt und ein breites Spektrum an deskriptiven Metadaten unterstützt, kann ein Manifest auch zu Zitationszwecken oder allgemein zur Verlinkung des damit verbundenen Korpus genutzt werden. Es fungiert somit als öffentlicher Platzhalter für die nicht direkt einsehbaren geschützten Korpusinhalte.

Beim folgenden Beispiel handelt es sich um ein Manifest mit Informationen zu Primärdaten im Repositorium, Rechteinhaber*innen und voreingestellten Werten für die statische Auszugsgenerierung. Das beschriebene Beispielkorpus ist eine 100-seitige PDF-Datei (verlinkt im `"xmp:primaryData"`-Block), bei der die ersten 10 Prozent im Falle von statischer Auszugsgenerierung geliefert werden sollen (spezifiziert im `"xmp:staticExcerpt"`-Block). Aus Platzgründen ist der `"xmp:manifests"`-Block für zusätzliche Korpus-Metadaten ohne Inhalt dargestellt:

```
{ "@type": "xmp:manifest", "@context": "http://www.uni-stuttgart.de/xsample/json-ld/manifest",
"xmp:description": "Plain manifest with no customization (first 10%)", "xmp:corpora":
[ { "@type": "xmp:corpus", "xmp:primaryData": { "@type": "xmp:dataverseFile", "xmp:segments":
100, "xmp:sourceType": "xmp:pdf", "xmp:id": 26 }, "xmp:legalNote": { "@type": "xmp:legalNote",
"xmp:author": "The XSample Team", "xmp:title": "XSample Test Corpus", "xmp:publisher": "The XSample
Project", "xmp:year": 2021 }, "xmp:description": "100 page test corpus", "xmp:id": "root" } ],
"xmp:staticExcerpt": { "@type": "xmp:span", "xmp:begin": 0, "xmp:end": 10, "xmp:spanType":
"xmp:relative" }, "xmp:manifests": []}
```

Sind alle Vorbereitungsschritte abgeschlossen, können Nachnutzer*innen über die Dataverse-Oberfläche (Abbildung 2) eine Zugriffsanfrage auf die XSample-Manifeste stellen. Wenn diese durch die Infrastrukturbetreiber*innen akzeptiert wird, dürfen die Nachnutzer*innen auf den XSample-Server (Abbildung 3) weitergeleitet werden, wo die eigentliche Konfiguration und Erstellung der Auszüge erfolgt. Da einzelnen Nutzer*innen nach § 60c UrhG jeweils nur maximal 15 Prozent eines geschützten Werkes ausgegeben werden dürfen und diese Obergrenze auch über wiederholte Anfragen hinweg eingehalten werden muss, bedarf es einer sehr genauen Protokollierung bereits ausgegebener Auszüge. Zu diesem Zweck werden die eindeutig identifizierbaren Dataverse-Accounts verwendet, was wiederum zur Folge hat, dass zur Nutzung des XSample-Services zwingend ein Account im jeweils verknüpften Dataverse-Repositorium notwendig ist und unregistrierte Dritte keinen Zugriff erhalten können.

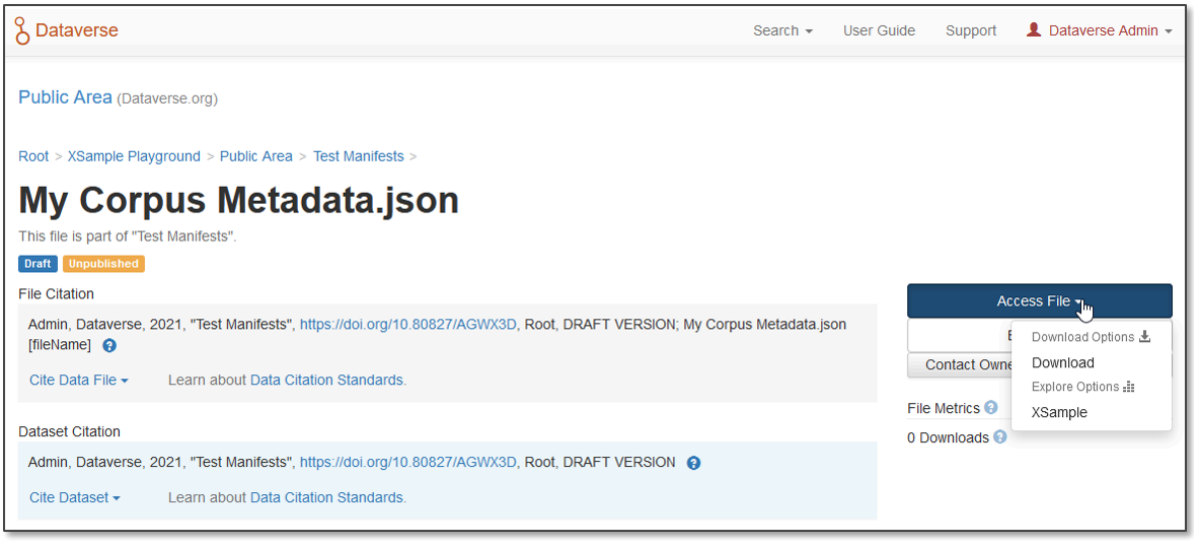

Abb. 2: Dataverse-Oberfläche für ein XSample-Manifest. Rechts unten kann die Weiterleitung auf den XSample-Server angestoßen werden. [Gärtner 2021]

3.4 Auswahl der Auszüge

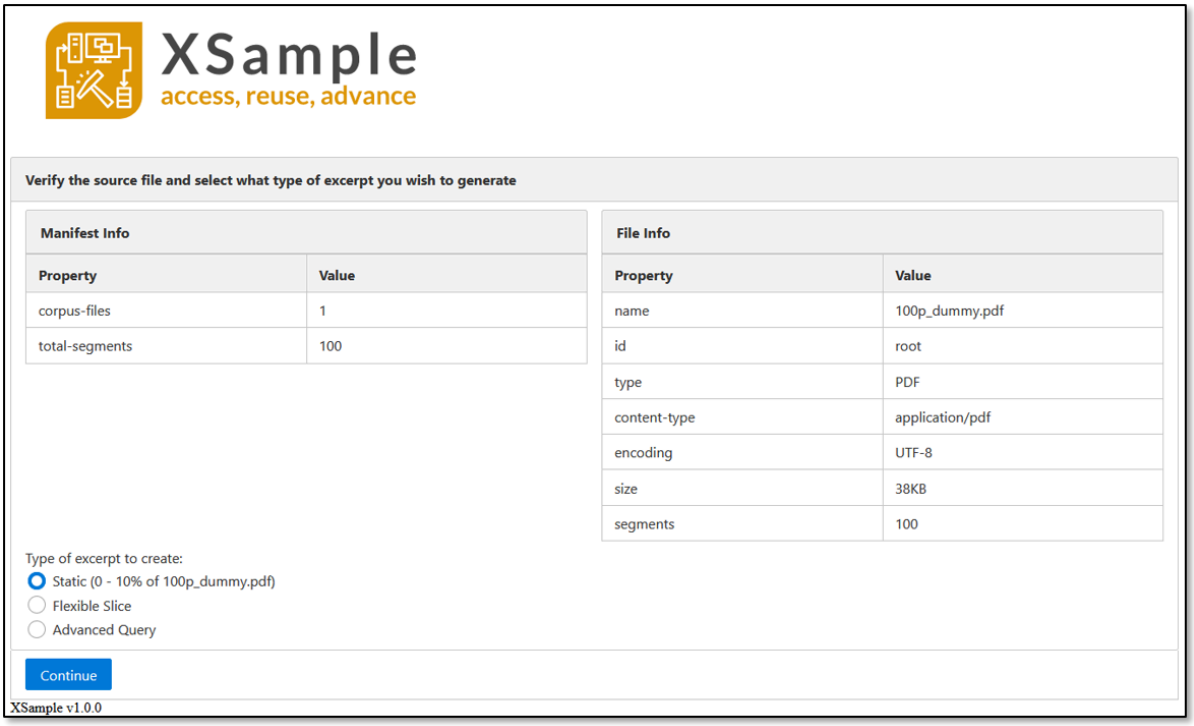

Abb. 3: Startseite des XSample-Servers nach Weiterleitung aus dem zugehörigen Dataverse und Validierung der Manifest-Datei. [Gärtner 2021]

Nach der Weiterleitung auf den XSample-Server erfolgt zunächst eine Validierung des Manifests auf formale Korrektheit und Verfügbarkeit der verlinkten (Korpus-)Ressourcen. Anschließend haben Nutzer*innen die Möglichkeit, zwischen drei Arten der Auszugsgenerierung (vgl. Abbildung 3, unten links) auszuwählen: Als simpelste Lösung kann ein statisch definierter Auszug (beispielsweise die ersten 10 Prozent oder ein anderer im Manifest definierter Abschnitt) gewählt werden. Wird mehr Kontrolle über die Zusammensetzung des Auszugs gewünscht, bietet die zweite Alternative (vgl. Abbildung 4) die Möglichkeit, die Auszugsgrenzen innerhalb der Ursprungsdaten frei zu definieren (z. B. die Seiten 20 bis 33). Auch hier wird eine

zusammenhängende Sequenz an Seiten / Segmenten geliefert. In jedem Falle stehen die Auszugsdaten am Ende des Workflows direkt als zip-Datei zum Download zur Verfügung, zusammen mit bibliografischen Informationen zu den Auszügen und den Rechteinhaber*innen.

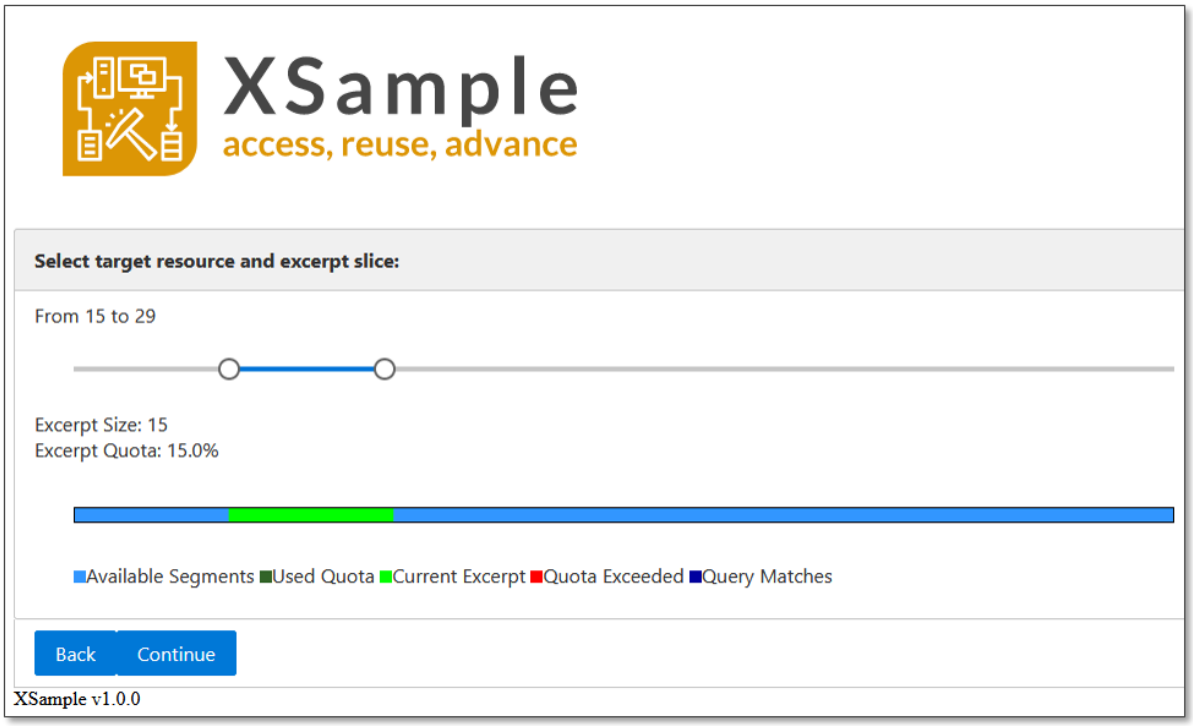

Abb. 4: Grafische Oberfläche zur flexiblen Auswahl der Auszugsgrenzen. [Gärtner 2021]

Die ersten beiden Verfahren der Auszugsgenerierung bieten den Nachnutzer*innen verhältnismäßig wenig Flexibilität bei der Auswahl der Textausschnitte, sodass sie für bestimmte Forschungsanliegen ungeeignet sind. Dies ist insbesondere dann der Fall, wenn sich die Nutzer*innen nur für sehr spezifische Phänomene oder Passagen interessieren. Um dem gerecht zu werden, wird als dritte Alternative eine Korpusanfrageschnittstelle[34] integriert, die Suchanfragen auf Basis der im Korpus enthaltenen Annotationen ermöglicht. Dadurch lassen sich beispielsweise gezielt bestimmte syntaktische Konstruktionen finden[35], die dann als Kandidaten für die Auszugserstellung genutzt werden (vgl. Abbildung 5). Basierend auf diesen Suchergebnissen und den Alignierungsinformationen werden die auszugebenden Segmente der Primärdaten (zumeist Seiten) bestimmt. Somit lässt sich sicherstellen, dass die Auszüge optimal auf die individuellen Bedürfnisse der Nutzer*innen zugeschnitten sind. Da Nutzer*innen zu diesem Zeitpunkt der Auszugsgenerierung noch kein Zugriff auf die geschützten Daten gewährt werden kann, wird lediglich eine visuelle Verteilung der Treffer und möglicher Auszugssegmente angeboten. Etablierte Such- und Visualisierungswerkzeuge wie ANNIS[36] oder KorAP[37] stellen zwar umfangreiche Such- und Exportmöglichkeiten zur Verfügung, bieten aber nicht diese notwendige Abschirmung der Daten bis zur finalen Auszugserstellung. Die den beiden erwähnten und anderen bestehenden Suchwerkzeugen zugrunde liegenden Anfragesprachen und -Schnittstellen könnten allerdings als Alternativen zu den im Prototypen integrierten Optionen für ICARUS und ICARUS2 dienen.

[34] Vgl. Gärtner 2020.
[35] Sofern entsprechende Annotationen vorliegen.
[36] Vgl. Krause / Zeldes 2016.
[37] Vgl. Diewald et al. 2106.

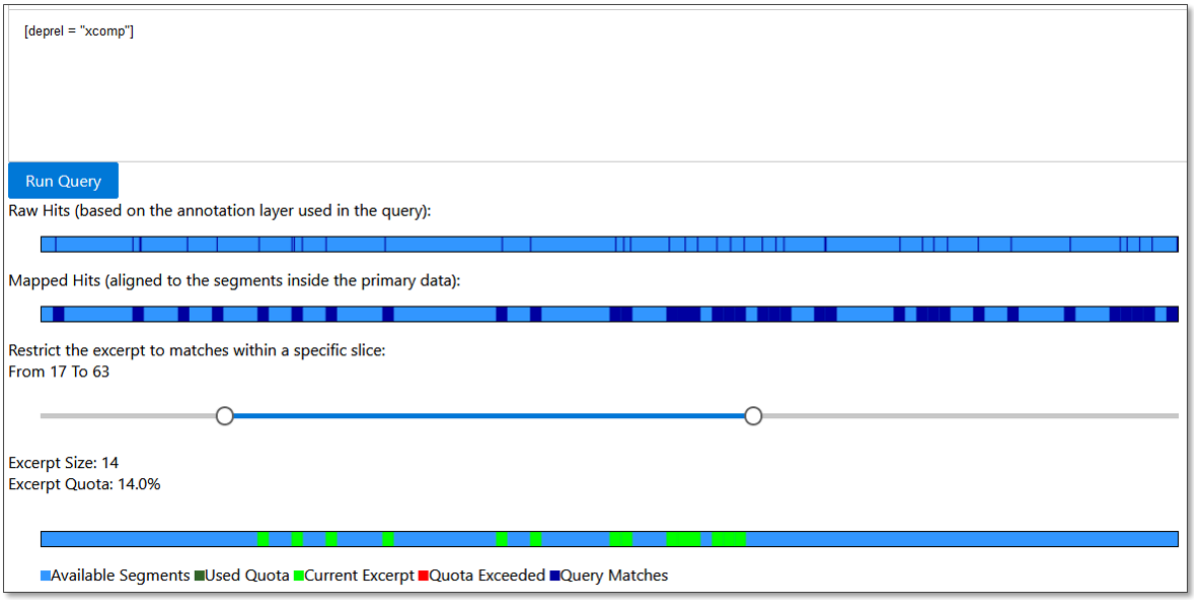

Abb. 5: Exemplarische Oberfläche zur Auszugsgenerierung mittels Suchanfrage basierend auf enthaltenen Annotationen. Die Verteilung der Suchergebnisse über das Korpus wird direkt visualisiert und Nutzer*innen anschließend die Möglichkeit gegeben, die für sie relevanten Segmente exakt auszuwählen. [Gärtner 2021]

3.5 Nachhaltigkeit

Um eine langfristige Nachnutzung des XSample-Servers zu gewährleisten, muss dieser dauerhaft an der Universitätsbibliothek Stuttgart (in ihrer Rolle als Forschungsinfrastruktureinrichtung) als Dienst zur Verfügung stehen. Wie in Kapitel 3.2 beschrieben, wird an der Universität Stuttgart ein Datenrepositorium auf Basis von Dataverse eingesetzt. Das Datenrepositorium der Universität Stuttgart (DaRUS) steht bereits als etablierter Dienst zur Verfügung. Mit dem XSample-Server zur Auszugsgenerierung kommt ein weiterer Dienst hinzu, der ebenfalls gewartet und weiterentwickelt werden muss. Bisher läuft dieser Dienst nicht im Produktivbetrieb, eine Verstetigung wird angestrebt. Dazu wird gerade im Forschungsdatenmanagement-Team ein Betriebskonzept für Code-Output von Forschungsprojekten erarbeitet. Während der Projektphase sollen die technischen Abhängigkeiten und notwendigen Informationen über den Code dokumentiert werden, darüber hinaus muss eine fachliche Ansprechperson benannt werden, die auch noch nach Projektende inhaltlich Auskunft geben kann. In einer zweiten Phase nach Projektende startet eine Bewährungsphase, in der evaluiert wird, ob der Dienst genutzt wird. In dieser Phase finden notwendige Wartungen statt. Damit soll verhindert werden, dass die Anwendung nach Projektende nicht mehr weiter nutzbar ist. Da aber nicht alle Dienste weiterentwickelt und gepflegt werden können, werden nur diejenigen weiterhin angeboten, die sich bewährt haben.

Die Software für den Prototypen ist open-source öffentlich verfügbar und kann somit auch von anderen Einrichtungen genutzt werden, um eine eigene Instanz des XSample-Servers zu betreiben. Neben einem laufenden Dataverse-Server wird lediglich eine SQL-basierte Datenbank benötigt, um den XSample-Server in bestehende Infrastruktur integrieren zu können. Die Anforderungen an Rechenleistung und Speicherplatz für den Prototypen sind hierbei überschaubar.

Zwar ist der aktuelle Prototyp auf die Schnittstelle zu Dataverse beschränkt, der XSample-Workflow lässt sich aber auf beliebige Repositoriensoftware mit ähnlichen Eigenschaften übertragen. Entsprechend kann mit mäßigem Aufwand auch der XSample-Server angepasst werden, um mit anderen Repositorien interagieren zu können. Etwas komplexer gestaltet sich die Unterstützung zusätzlicher Formate, da hierbei sichergestellt werden muss, dass sowohl die Implementierung des XSample-Servers selbst als auch die Korpusanfrageschnittstelle im Hintergrund ein gegebenes Format lesen können. Im Falle der Auszugskomponente im XSample-Server kommt noch das Schreiben der im Auszug enthaltenen Daten im entsprechenden Format hinzu.

4. Nutzungsszenarien

Der Bedarfsermittlung und Erprobung der XSample-Infrastruktur dienen zwei Anwendungsfälle aus Linguistik und Literaturwissenschaft, anhand derer die konkrete Umsetzung vor dem Hintergrund möglicher (Nach-)Nutzungsszenarien veranschaulicht wird. Der Fokus liegt dabei auf den jeweiligen Vor- und Nachteilen des Auszugsverfahrens im Vergleich zum Prinzip der abgeleiteten Textformate.

4.1 Erstes Nutzungsszenario: Wissenschaftssprache

Das erste Nutzungsszenario beschäftigt sich mit den Wissenschaftssprachen der Disziplinen Literaturwissenschaft, Linguistik und Philosophie. Es handelt sich dabei um die Replikation einer Studie[38] zur Frage, wie sich die Wissenschaftssprachen von Linguistik und Literaturwissenschaft voneinander unterscheiden. Als Datengrundlage der Originalstudie dienen jeweils 30 Dissertationen aus den beiden Fächern. Die Unterschiede zwischen den beiden Teilkorpora werden in einem datengeleiteten Verfahren mithilfe von maschinellem Lernen ermittelt. Daran anschließend erfolgt eine Interpretation der deutlichsten Unterschiede vor dem Hintergrund wissenschaftstheoretischer Merkmale der beiden Disziplinen. In der Replikationsstudie werden im Wesentlichen zwei Modifikationen vorgenommen:[39] Erstens erfolgt eine Veränderung der Variable ›Textsorte‹ von Dissertationen hin zu Zeitschriftenartikeln. Während die Autor*innen von Dissertationen mit ihren Texten zwar ihre Beherrschung der fachtypischen Wissenschaftssprache nachweisen, durchlaufen Zeitschriftenartikel in der Regel mehr Schritte der Qualitätssicherung. Sie werden außerdem von der Fachgemeinschaft breiter rezipiert, sodass sie auch als repräsentativer für die jeweilige fachspezifische Wissenschaftssprache gelten können. Zweitens wird die Datengrundlage um die Philosophie als drittes geisteswissenschaftliches Fach erweitert. Dadurch werden Literaturwissenschaft und Linguistik nochmals aus einer neuen Perspektive beleuchtet, nämlich im Kontrast zur Philosophie. Außerdem ist mit der Hinzunahme des dritten Faches ein Schritt dahingehend getan, Aussagen über die Wissenschaftssprache der Geisteswissenschaften im Allgemeinen zu treffen.

Das Korpus umfasst insgesamt 135 Zeitschriftenartikel, jeweils 45 pro Fach. Alle Texte werden automatisch mit Lemmata, Wortarten und syntaktischen Dependenzen annotiert. Für den XSample-Workflow werden die annotierten Texte im CoNLL-2009-Format an der Universitätsbibliothek Stuttgart gespeichert. Die Ursprungsdaten im PDF-Format werden gemeinsam mit den Annotationen im Repositorium aufbewahrt. Um sicherzustellen, dass von den annotierten Daten wieder auf die PDF-Seiten der Ursprungsdaten geschlossen werden kann, müssen hierbei Informationen zur Alignierung der beiden Dateiformate gespeichert werden. Im Falle dieses Nutzungsszenarios erfolgt dies in Form einer einfachen Tabelle für jedes Dokument, die für jeden fortlaufend nummerierten Satz verzeichnet, auf welcher Seite oder welchen Seiten im PDF-Dokument er zu finden ist.

Im Fokus der Replikationsstudie steht der Teil der vorausgehende Studie[40], der Einzelwörter und Wortartentags (*Unigramme*) betrachtet.[41] In methodischer Hinsicht orientiert sich die Replikationsstudie eng am Original: Im ersten Schritt werden die Merkmale mit den größten Unterschieden zwischen den Teilkorpora auf datengeleitete Weise ermittelt. Zu diesem Zweck wird mit dem maschinellen Lernverfahren der *Support-Vector-Machine* (SVM) ein Klassifikator trainiert, der jeweils zwischen Texten aus zwei der drei Disziplinen unterscheiden soll. Die lineare SVM bietet die Möglichkeit, auf die Koeffizienten zuzugreifen, die für jedes Merkmal ausdrücken, wie hilfreich es für die Klassifikationsaufgabe war. Anhand dieser Koeffizienten wird ein Feature-Ranking wie in Tabelle 1 erstellt, das die größten Unterschiede zwischen den Teilkorpora (im Sinne der SVM) darstellt. Der zweite Schritt der Analyse besteht dann in der Interpretation dieses Feature-Rankings. Welche sprachlichen Unterschiede verbergen sich hinter den Merkmalen und (wie) können sie anhand wissenschaftstheoretischer Merkmale der Disziplinen erklärt werden? Während der erste Schritt rein datengeleitet arbeitet, muss für die Interpretation auf unterschiedliche Ressourcen zurückgegriffen werden: Erstens ist der erneute Rückgriff auf das Korpus notwendig, um verstehen zu können, wie das Merkmal in den Texten verwendet wird. Zweitens muss Wissen über die wissenschaftstheoretischen Merkmale der Disziplinen sowie bereits vorhandene empirische Erkenntnisse zu den vorliegenden Phänomenen herangezogen werden, um die Daten in einen funktionalen Zusammenhang setzen und gegebenenfalls erklären zu können.

[38] Bei der Originalstudie handelt es sich um Andresen 2022.
[39] Es handelt sich dementsprechend um eine sogenannte ›approximative Replikation‹ (Porte 2012, S. 8).
[40] Vgl. Andresen 2022.
[41] Die ursprüngliche Studie (Andresen 2022) legt zusätzlich einen besonderen Schwerpunkt auf Sequenzen, die entlang der syntaktischen Dependenzen im Satz gebildet werden.

Rang	Philosophie	Score	Literaturwissenschaft
1		-77,07	ADJA
2	PPER	38,19	
3		-34,98	NE
4		-33,23	VVFIN
5	PDAT	29,00	
6	FM	20,20	
7	VAFIN	17,55	
8	KON	14,84	
9	PDS	12,39	
10	PROAV	12,10	
11		-11,65	PRF
12	PPOSAT	11,18	
13		-10,73	ART
14		-10,16	VVPP
15		-8,10	VVINF

Tab. 1: Die distinktivsten Wortarten für die Unterscheidung von Philosophie und Literaturwissenschaft im Sinne der SVM. Das verwendete Tagset ist das STTS (Schiller et al. 1999). [Andresen 2022]

Diese Schritte werden im Folgenden am Beispiel der Analyse der Verwendung des Pronomens wir in den drei Disziplinen veranschaulicht. Im Zuge dessen wird auch diskutiert, welche Daten zur Überprüfung und Re-Validierung der Analyseresultate vonnöten sind.

Tabelle 1 zeigt das Ranking der 15 distinktivsten Wortarten für den Vergleich von Philosophie und Literaturwissenschaft. Während sich aus diesen Ergebnissen zahlreiche relevante Rückschlüsse auf die Unterschiede zwischen den Wissenschaftssprachen der beiden Disziplinen ziehen lassen, wird hier nur ein Merkmal in den Fokus genommen: Im Vergleich mit der Literaturwissenschaft zeichnet sich die Philosophie durch eine hohe Verwendungsfrequenz von Personalpronomen (PPER) aus. Der zusätzliche Rückgriff auf die *Token*-Ebene zeigt, dass dieser Unterschied insbesondere durch die Pronomen wir und es erzeugt wird. Dies wird hier zum Anlass genommen, die fachspezifische (bzw. gegebenenfalls auch zeitschriften-spezifische) wir-Verwendung differenzierter zu betrachten. Zu diesem Zwecke wurde aus jedem Korpus eine Stichprobe von 100 Sätzen, in denen wir verwendet wird, nach dem Zufallsprinzip ausgewählt und manuell in Bezug auf ihre Funktion klassifiziert. Dabei wurde auf die Klassifikation von wir-Verwendungen im deutschsprachigen akademischen Diskurs von Kresta[42] zurückgegriffen. Kresta unterscheidet vier Gebrauchsweisen von wir in deutschsprachigen akademischen Texten: Das Pronomen wird verwendet,

- a.) um auf die tatsächlichen Verfasser*innen eines Textes zu verweisen (Autor*innen-wir),
- b.) um ein Kollektiv, bestehend aus Verfasser*innen und Leser*innen eines Textes, zu bezeichnen (Teamwork-wir),
- c.) zur Bezeichnung fachspezifischer Kollektiva aus akademischen Verfasser*innen und Leser*innen (Fachkreis-wir) sowie
- d.) um auf alle Menschen zu referieren (Gemeinschafts-wir).

Die Stichproben zeigen (vgl. Tabelle 2) in den Texten fachspezifische Muster der wir-Verwendungen: So sind sich die literaturwissenschaftlichen und philosophischen Texte darin ähnlich, dass in beiden Gruppen die Verwendung des Gemeinschafts-wir dominiert, während in den linguistischen Aufsätzen die Verwendung des Autor*innen-wir vorherrscht. Die Dominanz des Autor*innen-wir in der Linguistik lässt sich dadurch erklären, dass die Texte tatsächlich mehrheitlich von mehreren Autor*innen verfasst wurden (siehe Beispiel 1). Diese Praxis scheint in der Linguistik weitaus üblicher zu sein als in den anderen beiden Fächern. Die Ähnlichkeit von Philosophie und Literaturwissenschaft in ihrer Verwendung des Gemeinschafts-wir mag hingegen verwundern, da die beiden Fächer häufig ihre formal-sprachlichen Unterschiede betonen. Insbesondere in der Philosophie dient die Kennzeichnung eines philosophischen Ansatzes als ›literarisch‹ oft der Kritik am philosophischen Gehalt desselben.[43] Eine Auswertung der konkreten Belegstellen zeigt jedoch, dass es sich in den beiden Fächern um unterschiedlich geartete Manifestationen des Gemeinschafts-wir handelt, die man wiederum mit landläufigen Kennzeichen der beiden Fächer

[42] Vgl. Kresta 1995, S.130–147, vgl. auch Steinhoff 2007, S. 206f.
[43] Vgl. zum Beispiel Jürgen Habermas' Kritik an der *Einebnung des Gattungsunterschiedes zwischen Philosophie und Literatur*, Habermas 1988, S. 217.

in Verbindung bringen kann: So dominiert in der Stichprobe aus der Philosophie ein Gebrauch des Gemeinschafts-wir, der letztendlich auf die grundlegenden Bedingungen des menschlichen Denkens und Handelns abzielt, wie das Beispiel 2 belegt. In der Stichprobe aus der Literaturwissenschaft, deren zentralen Tätigkeiten die Lektüre und Interpretation von Texten sind, wird in 31 der 54 Verwendungen des Gemeinschafts-wir genau auf jene Praxis verwiesen, indem eine Art ›ideale*r Leser*in‹ konstituiert wird (siehe Beispiel 3), weswegen in diesem Fall auch vom Leser*innen-wir gesprochen werden kann.

	Linguistik	Philosophie	Literaturwissenschaft
Autor*innen-wir	58	7	11
Teamwork-wir	32	31	16
Fachkreis-wir	3	10	17
Gemeinschafts-wir	7	48	54
davon: Leser*innen-wir			(31)
Sonstige	–	4	2

Tab. 2: Manuelle Kategorisierung der wir-Verwendung in einer Stichprobe von 100 Instanzen pro Disziplin. [Pichler 2022]

Folgende Textausschnitte sollen als Beispiele für die nach Fachrichtung unterschiedlichen wir-Verwendungen dienen:

1. Im Folgenden werden **wir** die Datengrundlage näher erläutern und anschließend kurz auf die von uns verwendeten korpuspragmatischen Analysewerkzeuge eingehen. (Lin_16)
2. Diese Fähigkeit wird im Gegenteil schrittweise erlernt bzw. angeeignet – so wie **wir** z. B. unsere Muttersprache lernen oder aneignen – nämlich durch einen Prozess der ›unbewussten induktiven Schlussfolgerung‹, die auf Regelmäßigkeiten oder Assoziationen unter unseren Sinneswahrnehmungen zurückzuführen ist. (Philo_33)
3. Zugespitzt könnte man sagen, dass der Begriff ›literarische Präsenz‹ ein Widerspruch in sich ist, weil die Erzählung zwar von den Präsenzerfahrungen ihrer Figuren erzählen kann, doch diese stets allein auf der Ebene der histoire ›präsent‹ sind, präsent also für den Erzähler – doch **wir**, die Leser, sind nicht der Erzähler; […] (Lit_03)

Vor dem Hintergrund möglicher Nachnutzungsszenarien lassen sich für das erste Nutzungsszenario verschiedene Datenbedarfe feststellen, die mit unterschiedlichen Phasen der Analyse verbunden sind: Für die datengeleitete Ermittlung distinktiver Merkmale ist es für Nachnutzer*innen oder Gutachter*innen ausreichend, wenn ihnen die Texte nur in Form von n-Gramm-Frequenzen, also in einem abgeleiteten Format, vorliegen. Auf der Grundlage von z. B. Wortartenfrequenzen kann dieser Teil der Analyse direkt reproduziert werden.[44] Außerdem ist es etwa möglich, die distinktiven Merkmale auf den gleichen Daten mithilfe anderer Verfahren zu ermitteln und methodische Vergleiche anzustellen. Naturgemäß wird die weiterführende Analyse auf genau solche Frequenzen eingeschränkt, die auch zur Verfügung gestellt werden. Eine flexible Anpassung der n-Gramme (etwa ihrer Länge oder der Art ihrer Generierung) ist nicht ohne weiteres möglich, im Großen und Ganzen werden die Bedarfe dieser Analysephase aber durch abgeleitete Textformate gedeckt.

Geht es hingegen um eine Interpretation der Daten, welche auf semantische und pragmatische Dimensionen abzielt, reichen Frequenzinformationen nicht mehr aus, um das Vorgehen in der Studie im Rahmen eines Gutachten zu bewerten oder eigene Schlüsse aus den Daten zu ziehen. Um konkrete (semantische oder pragmatische) Phänomene, wie zum Beispiel in Hinblick auf den Gebrauch von wir, zu verstehen und gegebenenfalls erklären zu können, ist es notwendig, konkrete Verwendungen im Korpus mitsamt ihrem Kontext zu sichten. Die notwendige Kontextgröße hängt dabei von der Natur des untersuchten Phänomens ab. Für die wir-Analyse wurden pro Fach 100 zufällige Sätze untersucht. Ein zumindest stichprobenartiger Zugriff auf Volltextdaten, wie er durch den XSample-Ansatz ermöglicht wird, ist zentral, um geisteswissenschaftlich fundierte Aussagen treffen und nachvollziehbar machen zu können.

[44] Der Schritt von den Originaldaten zu den Frequenzdaten kann weder auf Grundlage dieser Daten noch basierend auf Auszügen überprüft werden. Das ist bedauerlich, da bereits in dieser Phase richtungsweisende Entscheidungen getroffen werden (Findet eine Lemmatisierung statt? Werden Stoppwörter ausgeschlossen? Werden bestimmte Teile der Originaltexte nicht einbezogen? etc.).

4.2 Zweites Nutzungsszenario: Unzuverlässiges Erzählen

Der zweite Anwendungsfall setzt sich mit dem Phänomen des unzuverlässigen Erzählens (genauer: mit faktenbezogener Unzuverlässigkeit) auseinander, das in einigen literarischen Erzählungen auftritt. Faktenbezogenes unzuverlässiges Erzählen liegt in einem fiktionalen Text dann vor, wenn die Erzählinstanz unzutreffende, zweifelhafte oder in relevanter Hinsicht unvollständige Aussagen über die Fakten oder Ereignisse der erzählten Welt tätigt.[45]

Im Gegensatz zum ersten Anwendungsfall handelt es sich beim zweiten nicht um eine Replikationsstudie, sondern um eine Pilotstudie zu einem kürzlich gestarteten, auf drei Jahre ausgelegten Forschungsprojekt (CAUTION), das der Untersuchung der Schluss- und Argumentationsprozesse bei der Identifikation unzuverlässigen Erzählens durch Literaturwissenschaftler*innen bzw. Leser*innen gewidmet ist. Unzuverlässiges Erzählen gilt in der Literaturwissenschaft einerseits als stark interpretationsabhängiges Phänomen,[46] andererseits listet die Forschung zahlreiche sprachliche Indikatoren, die auf unzuverlässiges Erzählen hinweisen können.[47] Leser*innen können solche Merkmale – unter Rückgriff auf allgemeines Weltwissen sowie literarische und literaturwissenschaftliche Kontexte – zum Anlass nehmen, der Erzählinstanz eines fiktionalen Textes Unzuverlässigkeit zuzuschreiben. In diesem Rahmen entwickeln sie eine inhaltsspezifizierende Interpretation[48] des Textes, d. h. sie bilden Annahmen darüber, was in der fiktiven Welt des Textes wahr und was falsch ist.

Um die Schluss- und Argumentationsprozesse bei der Feststellung bzw. Zuschreibung unzuverlässigen Erzählens systematisch untersuchen zu können, sind in einer ersten Annäherung folgende Teilfragen relevant:

1. Welche Erzähler*innen bzw. Figuren treten in einer Erzählung auf?
2. Welche Äußerungen über die fiktive Welt der Erzählung treffen diese Instanzen?
3. Welche dieser Äußerungen betreffen Propositionen, deren Zutreffen in der fiktiven Welt in Frage steht?
4. Wie positionieren sich die relevanten Instanzen zu diesen Propositionen?
5. Gibt es textuelle Hinweise auf die Vertrauens(un)würdigkeit der relevanten Instanzen?

Zur Beantwortung dieser Fragen wird im Rahmen des zweiten Nutzungsszenarios explorativ-heuristisch eine Mischung aus automatisierten Text-Mining-Verfahren und manueller Annotation auf ein Testkorpus aus vier kurzen bis mittellangen Erzählungen und vier langen Erzählungen aus dem 19. bis 21. Jahrhundert angewandt. Für die Teilfragen (1) und (2) werden automatische Verfahren zur Erkennung von Named Entities[49] und *Redewiedergabe* [50] verwendet, zusätzlich wurden Koreferenzen exemplarisch manuell annotiert. Für Fragen (3) und (4) muss – wie es bei der Untersuchung genuin literaturwissenschaftlicher Konzepte oft notwendig ist – zunächst ein eigenes Annotationsschema entwickelt werden, das dann im Rahmen manueller Annotation auf die Texte angewandt wird.[51] Für Frage (5) wird exemplarisch eine Indikatorengruppe aus der Unzuverlässigkeitsforschung in den Fokus genommen: die Verwendung emotionaler bzw. wertender Sprache, für deren Erkennung eine Kombination aus automatischer *Sentimentanalyse* [52] und manueller *Emotionsanalyse* eingesetzt wird.

Für den Einsatz computergestützter Verfahren in der Literaturwissenschaft ist oft erheblicher Entwicklungsaufwand notwendig, sowohl konzeptionell im Rahmen der Operationalisierung literaturwissenschaftlicher Forschungsfragen als auch technisch im Hinblick auf die Anpassung oder Neuentwicklung von Tools.[53] Deswegen kann es bei der Auswertung der Pilotstudie noch nicht darum gehen, die übergeordnete Forschungsfrage zu den Schluss- und Argumentationsprozessen bei der Feststellung unzuverlässigen Erzählens zu beantworten. Dennoch kann ein Einblick in die Daten bereits in diesem Zwischenstadium aufschlussreich sein. Generell ist Forschung im Bereich der Digital Humanities (und besonders im Bereich der *Computational Literary Studies*) stärker als in den traditionellen Geisteswissenschaften durch »Prozessualität, Vorläufigkeit und ›Nichtwissen‹« [54] gekennzeichnet. Dies lässt sich durchaus als Stärke dieser Ansätze verstehen, weil dadurch die Zwischenschritte der Forschung und Entwicklung dokumentiert (und damit durch Dritte einsehbar) werden, die in nicht-digitalen literaturwissenschaftlichen Zugängen oft implizit bleiben.

[45] Vgl. Martínez / Scheffel 2009, S. 100; Kindt 2008, S. 48.
[46] Vgl. Yacobi 1981; Nünning 1999.
[47] Vgl. Nünning 1998; Allrath 1998.
[48] Vgl. Folde 2015, S. 366.
[49] Verwendet wurde hier der Stanford Named Entity Recognizer.
[50] Für die Erkennung von direkter Rede wurde ein simpler Tagger entwickelt, der auf der Identifikation von Anführungszeichen basiert; indirekte Rede wurde mithilfe eines verfügbaren Taggers annotiert. Alle erzeugten Annotationen wurden anschließend gesichtet und gegebenenfalls korrigiert.
[51] Für die manuelle Annotation wurde die Annotations- und Analyseumgebung CATMA verwendet.
[52] Zum Einsatz kam hier SentText, vgl. Schmidt et al. 2021.
[53] Vgl. Gius 2019; Pichler / Reiter 2021.
[54] Schruhl 2018.

Beim zweiten Nutzungsszenario dient eine Einsicht in die Daten durch Dritte also hauptsächlich dem Zweck, einen Einblick in den Operationalisierungsprozess der übergeordneten literaturwissenschaftlichen Fragestellung zu erhalten, oder ist dem Interesse an bestimmten Einzelphänomenen (etwa dem Sentiment) geschuldet. Nachnutzer*innen können beispielsweise prüfen, inwieweit die eingesetzten Text-Mining-Verfahren bereits für die Anwendung auf literarischen Texten adäquat sind oder ob die für die manuelle Annotation entwickelten Annotationsschemata geeignet sind, die im Fokus stehenden literarischen Phänomene zu fassen.

Für die Form, in der die Textdaten Dritten zugänglich gemacht werden sollten, bedeutet das im vorliegenden Zusammenhang Folgendes:

1. *Named Entity Recognition*: Inwieweit die *Named Entity Recognition* (mit zu diesem Zeitpunkt noch nicht eigens für das Korpus trainierten Modellen) auf literarischen Texten zu brauchbaren Ergebnissen führt, kann unter Umständen noch teilweise mittels abgeleiteter Textformate (z. B. bestimmter tokenbasierter Formate[55]) geprüft werden. Zusätzlich können Nachnutzer*innen sich damit ebenfalls einen ersten Eindruck hinsichtlich der im jeweiligen Text auftretenden Figuren verschaffen.[56]
2. *Automatische Sentimentanalyse*: Um zu beurteilen, ob durch das lexikonbasierte Vorgehen der automatischen Sentimentanalyse einzelne Wörter falsch klassifiziert wurden, ist der Rekurs auf den textuellen Kontext notwendig, den abgeleitete Textformate nicht ermöglichen. Ein erster Einblick in die Ergebnisse der automatischen Sentimenanalyse ist allerdings noch mit abgeleiteten Textformaten möglich, sofern für Nachnutzer*innen interessant ist, ob ein Text bzw. Korpus eher von negativen oder positiven Wörtern geprägt ist oder welche Wörter bzw. Wortfelder mit positiver oder negativer Polarität vorherrschen.
3. *Redewiedergabeerkennung*: Für die Prüfung der automatisch generierten Redewiedergabe-Annotationen sind abgeleitete Textformate ebenfalls nicht funktional, da die Annotationen längere Passagen betreffen bzw. ihre Korrektheit (insbesondere im Fall indirekter Rede) nur unter Rückgriff auf die fraglichen Textpassagen beurteilt werden kann. Auch der für die weitere Bearbeitung der übergeordneten Forschungsfrage ausschlaggebende Inhalt der Figurenrede kann nur durch Konsultation zusammenhängender Textpassagen untersucht werden.
4. *Koreferenzauflösung*: Bei den manuellen Annotationen zur Koreferenzauflösung sind sowohl zur Prüfung der Korrektheit als auch für die Bearbeitung der inhaltlichen Fragestellung (›Wer sagt was?‹) textuelle Kontexte notwendig.
5. *Manuelle Emotionsanalyse*: Für die manuelle Emotionsanalyse wurde im Rahmen des Nutzungsszenarios ein eigenes Tagset entwickelt, das auf den sprachlichen Indikatoren basiert, die in der erzähltheoretischen Forschungsliteratur als Hinweise auf die Emotionalität von Erzähler*innen (und damit auf ihre mögliche Unzuverlässigkeit) identifiziert werden. Derartige Operationalisierungen literaturwissenschaftlicher Forschungsfragen für die computergestützte Analyse sind oft langwierige Prozesse und benötigen im Rahmen von manueller (und meist kollaborativer) Annotation häufig mehrere Durchläufe.[57] Im Rahmen des zweiten Nutzungsszenarios hat ein erster dieser Durchläufe stattgefunden, in dem Forscher*innen bzw. Datenlieferant*innen einige Entscheidungen treffen mussten, die für den Nachvollzug der Ergebnisse durch Dritte relevant sein können. Hierfür ist nicht nur ein Einblick in die Annotationsschemata und Anwendungsrichtlinien notwendig, sondern ebenso der Zugriff auf zusammenhängende Textteile, die den textuellen Kontext zeigen und damit individuelle Annotationsentscheidungen potenziell nachvollziehbar machen. Einen ersten Eindruck von den Ergebnissen der manuellen Emotionsanalyse können sich Nachnutzer*innen – analog zur verwandten Sentimentanalyse – allerdings auch auf der Basis abgeleiteter Textformate verschaffen.
6. *Wahre Propositionen der erzählten Welt*: Den komplexesten und zugleich wichtigsten Operationalisierungs- und Annotationsschritt stellt im Rahmen des zweiten Nutzungsszenarios die manuelle Annotation der Sätze dar, die diejenigen Propositionen betreffen, deren Wahrheit in der fiktiven Welt eines Textes in Frage stehen. Genau wie im Fall der manuellen Emotionsanalyse müssen die Annotationskategorien erst in mehreren Durchläufen entwickelt werden, von denen der erste im Rahmen des vorliegenden Nutzungsszenarios stattfindet. Allerdings muss für die Entwicklung von Annotationsschemata und Guidelines hier noch mehr Vorarbeit geleistet werden als bei der Emotionsanalyse, da in der Unzuverlässigkeitsforschung für letztere bereits Listen mit textuellen Indikatoren zur Verfügung stehen, die vergleichsweise direkt in Annotationsschemata übertragen werden können. Die Annotation in Frage stehender Sätze erfordert dagegen grundsätzliche konzeptionelle und praktische Entscheidungen. Im Rahmen dieses Nutzungsszenarios wurden beispielsweise zunächst jeweils textspezifische Kategorien entwickelt, d. h. dass für jeden Text ca. zwölf zentrale, in Frage stehende Propositionen identifiziert und als Tagset umgesetzt wurden (z. B. für E. T. A. Hoffmanns *Der Sandmann* die Propositionen »Der dämonische Sandmann existiert«, »Der Sandmann will Nathanaels Leben zerstören«, »Advokat Coppelius und Wetterglashändler Coppola sind dieselbe Person« etc.). Mithilfe dieser spezifischen Tagsets wurden in den Texten jeweils Sätze annotiert, in denen die fraglichen Propositionen thematisiert werden, um überhaupt erst einmal die

[55] Vgl. Schöch et al. 2020.
[56] Tatsächlich ist (auch eine eigens trainierte) Named Entity Recognition nur in eingeschränktem Maße für die Identifikation der relevanten Akteur*innen / Instanzen geeignet, da gerade in potenziell unzuverlässigen Erzählungen häufig homodiegetische Erzähler*innen (›Ich-Erzähler*innen‹) auftreten, auf die nicht oder selten mit Eigennamen referiert wird.
[57] Vgl. Gius / Jacke 2017; Reiter 2020.

grundsätzliche Umsetzbarkeit des Ansatzes zu testen. Perspektivisch ist die Entwicklung eines generischen Tagsets für diese Annotationsaufgabe wünschenswert, in dem beispielsweise Propositionen in Typen (z. B. in *singular*, *particularized* und *general*) unterteilt und verschiedene Rollen der Propositionen im Rahmen von Argumenten (Prämisse und Konklusion) unterschieden werden. Deshalb ist zu erwarten, dass sich die Annotationskategorien und Anwendungsrichtlinien noch erheblich verändern werden. Aus diesem Grund kann der Nachvollzug der einzelnen Operationalisierungsschritte durch Dritte von besonderer Bedeutung sein. Dieser wird wieder durch Zugriff auf Annotationskategorien / Anwendungsrichtlinien im Manifest sowie (mindestens) zusammenhängende Textabschnitte ermöglicht. An dieser Stelle tritt ein weiterer Unterschied zur manuellen Emotionsanalyse zutage: Während bei der Emotionsanalyse anhand sprachlicher Indikatoren vornehmlich einzelne Wörter annotiert werden, betrifft die Annotation bei der Analyse in Frage stehender Propositionen mindestens Teilsätze. Tokenbasierte abgeleitete Textformate sind deswegen weder für den Nachvollzug der Kategorienentwicklung oder der Annotationsentscheidungen noch für einen ersten Einblick in die Ergebnisse der Annotation bzw. Analyse funktional.

Zusammenfassend lässt sich also festhalten, dass abgeleitete Textformate für den Nachvollzug der Operationalisierung literaturwissenschaftlicher Forschungsfragen, die in der durch das zweite Nutzungsszenario abgebildeten Forschungsphase im Vordergrund steht, nur schwer oder gar nicht verwendbar sind. Ein Auszugskonzept kann dagegen vielen der genannten Anforderungen begegnen.

Wie eingangs deutlich gemacht wurde, dient das zweite Nutzungsszenario als Pilotstudie zu einem umfangreicheren Projekt, das die Schluss- und Argumentationsprozesse bei der Feststellung bzw. Zuschreibung unzuverlässigen Erzählens untersucht. In diesem Zusammenhang wird für Nachnutzer*innen die Notwendigkeit, auf zusammenhängende Textteile zugreifen zu können, noch stärker in den Vordergrund rücken, denn unzuverlässiges Erzählen gilt als Phänomen, dessen Feststellung sich aus dem Zusammenspiel verschiedener sprachlicher Indikatoren, über den Text verteilter Informationen und Kontextinformationen ergibt. Aus diesem Grund bleibt letztlich noch zu prüfen, inwieweit (d. h. bei welchen Phänomenvarianten oder Texten) der Zugriff auf Textauszüge für Nachnutzer*innen ausreichend ist. Fest steht allerdings, dass sich ein Auszugsmodell, wie es in XSample entwickelt wurde, den Bedarfen dieser Anwendungsfälle deutlich stärker annähert als abgeleitete Textformate und Forschenden eine Möglichkeit bietet, (auch) an urheberrechtlich geschützten Texten solch komplexe literarische Phänomene zu untersuchen.

5. Fazit

Um urheberrechtlichen Einschränkungen bei der Auswahl, Verbreitung und Nachnutzung von Forschung von vornherein aus dem Weg zu gehen, konzentrieren viele digitale Geisteswissenschaftler*innen ihre korpusorientierte Forschung auf gemeinfreie Texte. In der Breite führt dies zu Verzerrungen der Forschungslandschaft, die inhaltlich wie methodologisch problematisch sind. Dieser Artikel ging daher von folgendem Desiderat für die Forschungscommunity in den digitalen Geisteswissenschaften aus: Der bestehende urheberrechtliche Rahmen sollte in der Praxis so gut es geht ausgenutzt werden, nicht zuletzt um die Relevanz von korpusorientierter Forschung auf geschützten Texten forschungspolitisch zu unterstreichen. Hier kommt einer forschungsgeleiteten Dateninfrastruktur die wichtige Rolle zu, Forscher*innen eine möglichst weitgehende, rechtskonforme Verwendung geschützter Texte zu ermöglichen.

Zwei Ansätze hierzu sind zum einen individuelle Lizenzvereinbarungen, zum anderen das jüngst vorgestellte Prinzip abgeleiteter Textformate. Der Austausch von Ergebnissen einer explorativen Forschungspraxis auf Basis von Fragestellungen, die für eine Interpretation die Einbeziehung relevanten Kontexts erforderlich machen, ist mit beiden Ansätzen aber nur sehr eingeschränkt möglich. Der vorliegende Beitrag schlägt daher eine infrastrukturelle Erweiterung des Instrumentariums vor, die auf der urheberrechtlich zulässigen Weitergabe von Textauszügen aufbaut. Um die Nützlichkeit dieses Ansatzes für das individuelle Forschungsvorhaben zu maximieren, ermöglicht der XSample-Workflow den Nutzer*innen, Textauszüge flexibel anhand von Suchanfragen an den Text und seine Annotationen auszuwählen.

Anhand zweier Nutzungsszenarien aus Sprach- und Literaturwissenschaft wurde beispielhaft gezeigt, welche Möglichkeiten und Grenzen sich aus dem Prinzip abgeleiteter Textformate sowie dem Auszugskonzept im Kontext konkreter geisteswissenschaftlicher Forschungsprojekte ergeben. Das linguistische Szenario vergleicht Wortfrequenzen zwischen drei Korpora, eine Aufgabe, die problemlos anhand von einfachen Frequenzlisten – d. h. auf Basis abgeleiteter Textformate – reproduziert werden kann. Jedoch erfordert der nächste Schritt, die Interpretation der quantitativen Befunde, eine Rekontextualisierung der Ergebnisse und damit Zugriff auf die zu untersuchenden Textstellen in ihrem Kontext. Hier stößt das Prinzip abgeleiteter Textformate an seine Grenzen; durch das Auszugskonzept kann dieser Schritt hingegen in einem für den Anwendungsfall ausreichendem Maße geleistet werden.

Das zweite Szenario beschäftigt sich mit dem Phänomen des unzuverlässigen Erzählens, bei dem sich schnell zeigt, dass abgeleitete Textformate nicht sinnvoll eingesetzt werden können. Das betrifft zum einen den Nachvollzug der Operationalisierung der literaturwissenschaftlichen Kategorien, für den der nähere textuelle Kontext notwendig ist, zum anderen die Analyse und Interpretation der Annotationen als Indikatoren für Vorkommnisse unzuverlässigen Erzählens, für die auch der weitere textuelle Kontext von Bedeutung ist. Lediglich Vorverarbeitungsschritte wie eine Named Entity Recognition können über abgeleitete Formate, etwa mittels Frequenzdaten, nachgenutzt werden. Der Zugriff auf individuell ausgewählte Textauszüge ist für den literaturwissenschaftlichen Anwendungsfall somit deutlich vielversprechender. Allerdings ist anzumerken, dass für gewisse Interpretationen auch der ganze Text vorliegen muss. Hier könnte das Auszugskonzept zu einer ersten Sichtung und Bewertung des Materials dienen, vor dem Hintergrund komplexer literaturwissenschaftlicher Fragestellung aber an seine Grenzen stoßen.

Die zwei Nutzungsszenarien können die Breite geisteswissenschaftlicher Fragestellungen und Methoden nur in begrenztem Maß abbilden. Sie machen jedoch deutlich, dass selbst Analysen, die einen quantitativen, auf automatisierte Verfahren bauenden Zugang zu den Daten nutzen, für die Interpretation am Ende auf Kontextinformationen angewiesen sind. Nur dieser Schritt macht die Analyse an die Geisteswissenschaften anschlussfähig und für die Forschungscommunity nachvollziehbar. Die Arbeit mit Frequenzdaten, zu denen keine Kontextinformationen zur Verfügung stehen, birgt auch die Gefahr, zu Interpretationen zu verleiten, die nicht durch die Daten gedeckt sind. Insofern erscheint es für viele Forschungsszenarien in den digitalen Geisteswissenschaften sinnvoll, eine Kombination beider Verfahren anzustreben.

Alle hier diskutierten Verfahren bleiben selbstverständlich Behelfslösungen. Aus Sicht der Forschung wäre die generelle Möglichkeit, Forschungsdaten einschließlich der zugrundeliegenden Texte für wissenschaftliche Zwecke uneingeschränkt zu teilen, das bei weitem produktivste Vorgehen. Die Interessen der Rechteinhaber*innen müssen dabei natürlich berücksichtigt werden. Gegebenenfalls müssten langfristig etwa die Richtlinien der Forschungsfinanzierung aus öffentlichen Quellen angepasst werden, um bei der Verwendung von urheberrechtlich geschützten Werken in berechtigten Fällen eine vorgelagerte Kompensation für eine langfristige Nachnutzung zu ermöglichen. In vielen Fällen ließe sich ein Interessenausgleich von Forschung und Rechteinhaber*innen erreichen. Langfristig ist zu hoffen, dass die Politik den rechtlichen Rahmen mit dieser Zielsetzung weiterentwickelt. Unter den aktuell gegebenen Umständen erlauben zum einen die Veröffentlichung von abgeleiteten Textformaten und zum anderen der gezielte Zugriff auf genau die Auszüge des Textes, die für eine gegebene Fragestellung relevant sind, eine zwar eingeschränkte, in vielen Fällen aber hinreichende Reproduktion und Nachnutzung urheberrechtlich geschützter Forschungsdaten.

Bibliografische Angaben

Gaby Allrath: »But why will you say that I am mad?« Textuelle Signale für die Ermittlung von unreliable narration. In: Unreliable Narration. Studien zur Theorie und Praxis unglaubwürdigen Erzählens in der englischsprachigen Erzählliteratur. Hg. von Ansgar Nünning / Carola Surkamp / Bruno Zerweck. Trier 1998, S. 59–80. [Nachweis im GVK]

Melanie Andresen: Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen. Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen. Tübingen 2022. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP), 10). [Nachweis im GVK]

Nils Diewald / Michael Hanl / Eliza Margaretha / Joachim Bingel / Marc Kupietz / Piotr Bański / Andreas Witt: KorAP Architecture. Diving in the Deep Sea of Corpus Data In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Hg. von European Language Resources Association (ELRA). (LREC 2016: Portorož, 23.–28.05.2016). Paris 2016: European Language Resources Association (ELRA), S. 3586–3591. PDF. [online] [Nachweis im GVK]

Thomas Dreier / Gernot Schulze: UrhG – Urheberrechtsgesetz, Verwertungsgesellschaftengesetz, Kunsturhebergesetz. Kommentar. 6. Auflage. München 2018. [Nachweis im GVK]

Thomas Dreier / Gernot Schulze: UrhG – Urheberrechtsgesetz, Urheberrechts-Diensteanbieter-Gesetz, Verwertungsgesellschaftengesetz, Nebenurheberrecht, Kunsturheberrecht. Kommentar. 7. Auflage. München 2022. [Nachweis im GVK]

Christian Folde: Grounding Interpretation. In: British Journal of Aesthetics 55 (2015), H. 3, S. 361–374. [Nachweis im GVK]

Deutsche Forschungsgemeinschaft: Leitlinien zur Sicherung der guten wissenschaftlichen Praxis, Kodex, 2019. DOI: 10.5281/zenodo.6472827

Markus Gärtner / Katrin Schweitzer / Kerstin Eckart / Jonas Kuhn: Multi-modal Visualization and Search for Text and Prosody Annotations. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: System Demonstrations. Hg. von Association for Computational Linguistics. (ACL 53: Beijing, 27.–29.07.2015). Red Hook, NY 2015, S. 25–30. PDF. DOI: 10.3115/v1/P15-4005

Markus Gärtner / Jonas Kuhn: A Lightweight Modeling Middleware for Corpus Processing. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. Hg. von European Language Resources Association (ELRA). (LREC 2018: Miyazaki, Mai 2018), Miyazaki 2018, S. 1087–1095. PDF. [online]

Markus Gärtner: The Corpus Query Middleware of Tomorrow – A Proposal for a Hybrid Corpus Query Architecture. In: Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora. Hg. von Piotr Bański / Adrien Barbaresi / Simon Clematide / Marc Kupietz / Harald Lüngen / Ines Pisetta. (CMLC 8, Marseille, 11.–16.05.2020) Stroudsburg, PA 2020, S. 31–39. [online]

Markus Gärtner / Felicitas Kleinkopf / Melanie Andresen / Sybille Hermann: Corpus Reusability and Copyright – Challenges and Opportunities. In: Proceedings of the Workshop on Challenges in the Management of Large Corpor. Hg. von Harald Lüngen / Marc Kupietz / Piotr Bański / Adrien Barbaresi / Simon Clematide / Ines Pisetta. (CMLC 9, Limerick, 12.07.2021) Mannheim 2021, S. 10–19. DOI: 10.14618/ids-pub-10467 [Nachweis im GVK]

Evelyn Gius: Computationelle Textanalysen als fünfdimensionales Problem: Ein Modell zur Beschreibung von Komplexität. In: LitLab Pamphlet 8 (2019). [online]

Evelyn Gius / Janina Jacke: The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis. In: International Journal of Humanities and Arts Computing 11 (2017), H. 2, S. 233–254. DOI: 10.3366/ijhac.2017.0194 [Nachweis im GVK]

Jürgen Habermas: Der philosophische Diskurs der Moderne. Zwölf Vorlesungen. Frankfurt / Main 1988. (= Suhrkamp-Taschenbuch Wissenschaft, 749). [Nachweis im GVK]

Matthew Lee Jockers: Macroanalysis: Digital methods and literary history. Urbana, IL u. a. 2013. [Nachweis im GVK]

Tom Kindt: Unzuverlässiges Erzählen und literarische Moderne: eine Untersuchung der Romane von Ernst Weiß. Tübingen 2008. (= Studien zur deutschen Literatur, 184). [Nachweis im GVK]

Felicitas Kleinkopf: Text- und Data-Mining. Die Anforderungen digitaler Forschungsmethoden an ein innovations- und wissenschaftsfreundliches Urheberrecht. (= Schriftenreihe des Archivs für Urheber- und Medienrecht, 300). Baden-Baden 2022. PDF. DOI: 10.5771/9783748935360

Felicitas Kleinkopf / Janina Jacke / Markus Gärtner: Text- und Data-Mining – Urheberrechtliche Grenzen der Nachnutzung wissenschaftlicher Korpora bei computergestützten Verfahren und digitalen Ressourcen. In: MMR. Zeitschrift für IT-Recht und Recht der Digitalisierung 24 (2021), H. 3, S. 196–200. DOI: 10.18419/opus-11445 [Nachweis im GVK]

Felicitas Kleinkopf / Thomas Pflüger: Digitale Bildung, Wissenschaft und Kultur – Welcher urheberrechtliche Reformbedarf verbleibt nach Umsetzung der DSM-RL durch das Gesetz zum Urheberrecht im digitalen Binnenmarkt? In: Zeitschrift für Urheber- und Medienrecht 56 (2021), H. 8 / 9, S. 643–655. [Nachweis im GVK]

Thomas Krause / Amir Zeldes: ANNIS3. A New Architecture for Generic Corpus Query and Visualization. In: Digital Scholarship in the Humanities 31 (2016). H. 1, S. 118–139. 24.10.2014. DOI: 10.1093/llc/fqu057

Ronald Kresta: Realisierungsformen der Interpersonalität in vier linguistischen Fachtextsorten des Englischen und des Deutschen (= Theorie und Vermittlung der Sprache, 24). Frankfurt / Main u. a. 1995. [Nachweis im GVK]

Matías Martínez / Michael Scheffel: Einführung in die Erzähltheorie. 8. Auflage. (= C.-H.-Beck-Studium). München 2009. [Nachweis im GVK]

Ansgar Nünning: »Unreliable Narration« zur Einführung. Grundzüge einer kognitiv-narratologischen Theorie und Analyse unglaubwürdigen Erzählens. In: Unreliable Narration. Studien zur Theorie und Praxis unglaubwürdigen Erzählens. Hg. von Ansgar Nünning / Bruno Zerweck / Carola Surkamp. Trier 1998, S. 3–39. [Nachweis im GVK]

Ansgar Nünning: Unreliable, Compared to What? Towards a Cognitive Theory of »Unreliable Narration«. Prolegomena and Hypotheses. In: Grenzüberschreitungen. Narratologie im Kontext / Transcending Boundaries. Narratology in Context. Hg. von Walter Grünzweig / Andreas Solbach. Tübingen 1999, S. 53–73. [Nachweis im GVK]

Axel Pichler / Nils Reiter: Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers hermeneutische Modellinterpretation von Kleists Das Erdbeben in Chili. In: Journal of Literary Theory 15 (2021), H. 1–2, S. 1–29. [online] [Nachweis im GVK]

Graeme Porte: Introduction. In: Replication Research in Applied Linguistics. Hg. von Graeme Porte. (= Cambridge Applied Linguistics Series). Cambridge u. a. 2012, S. 1–17. [Nachweis im GVK]

Benjamin Raue: Die Freistellung von Datenanalysen durch die neuen Text und Data Mining-Schranken. In: Zeitschrift für Urheber- und Medienrecht 56 (2021), H. 10, S. 793-802. [Nachweis im GVK]

Nils Reiter: Anleitung zur Erstellung von Annotationsrichtlinien. In: Reflektierte algorithmische Textanalyse. Hg. von Nils Reiter / Axel Pichler / Jonas Kuhn. Berlin u. a. 2020, S. 193–202. DOI: 10.1515/9783110693973-009 [Nachweis im GVK]

Richtlinie (EU) 2019/790 des Europäischen Parlaments und des Rates vom 17. April 2019 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt und zur Änderung der Richtlinien 96/9/EG und 2001/29/EG. [online]

Anne Schiller / Simone Teufel / Christine Thielen / Christine Stöckert: Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). 1999. PDF: [online]

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 5 (2020). DOI: 10.17175/2020_006

Urheberrecht. UrhG, KUG, VGG. Kommentar. Hg. von Gerhard Schricker / Ulrich Loewenheim / Matthias Leistner. 6. neu bearbeitete Auflage. München 2020. [Nachweis im GVK]

Friederike Schruhl: Objektumgangsnormen in der Literaturwissenschaft. In: Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden. Hg. von Martin Huber / Sybille Krämer. Wolfenbüttel 2018. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 3) DOI: 10.17175/sb003_012

Thomas Schmidt / Johanna Dangel / Christian Wolff: SentText: A Tool for Lexicon-based Sentiment Analysis in Digital Humanities. In: Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science. Hg. von Christian Wolff / Thomas Schmidt. (ISI 16, Regensburg, 08–10.03.202) Glückstadt 2021, S. 156–172. DOI: 10.5283/epub.44943 [Nachweis im GVK]

Torsten Steinhoff: Wissenschaftliche Textkompetenz: Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten. Tübingen 2007. (= Reihe Germanistische Linguistik, 280) [Nachweis im GVK]

Mark D. Wilkinson / Michel Dumontier / IJsbrand Jan Aalbersberg / Gabrielle Appleton / Myles Axton / Arie Baak / Niklas Blomberg / Jan-Willem Boiten / Luiz Bonino da Silva Santos / Philip E. Bourne / Jildau Bouwman / Anthony J. Brookes / Tim Clark / Mercè Crosas / Ingrid Dillo / Olivier Dumon / Scott Edmunds / Chris T. Evelo / Richard Finkers / Alejandra Gonzalez-Beltran / Alasdair J.G. Gray / Paul Groth / Carole Goble / Jeffrey S. Grethe / Jaap Heringa / Peter A.C 't Hoen / Rob Hooft / Tobias Kuhn / Ruben Kok / Joost Kok / Scott J. Lusher / Maryann E. Martone / Albert Mons / Abel L. Packer / Bengt Persson / Philippe Rocca-Serra / Marco Roos / Rene van Schaik / Susanna-Assunta Sansone / Erik Schultes / Thierry Sengstag / Ted Slater / George Strawn / Morris A. Swertz / Mark Thompson / Johan van der Lei / Erik van Mulligen / Jan Velterop / Andra Waagmeester / Peter Wittenburg / Katherine Wolstencroft / Jun Zhao / Barend Mons: The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific Data 3 (2016), Artikelnummer 160018. DOI: 10.1038/sdata.2016.18 [Nachweis im GVK]

Tamar Yacobi: Fictional reliability as a communicative problem. In: Poetics Today 2 (1981), H. 2, S. 113–126. [Nachweis im GVK]

Abbildungs- und Tabellenverzeichnis