IRDT PAPER*SERIES*

IRDT Institut für Recht und Digitalisierung Trier

IRDT PAPER*SERIES* Nr. 5

Verfahrensschritte bei dem Einsatz von Text und Data Mining-Verfahren in den Geisteswissenschaften

Katharina Erler-Fridgen[1]

Version 1.0 (02.06.2022), CC BY-SA 4.0.

Werden urheberrechtlich geschützte[2] Texte mit Hilfe von Text und Data Mining Verfahren analysiert, unterliegen die vorgenommenen Handlungen grundsätzlich urheberrechtlichen Restriktionen. Denn werden Vervielfältigungen oder Entnahmen aus Datenbanken[3] vorgenommen oder der Analyse zugrundeliegende Text(teil)e präsentiert[4], können die Verwertungsrechte des Urhebers beeinträchtigt werden. Ist dies der Fall, bedarf es einer urheberrechtlichen Gestattung: Die Text und Data Mining Schranken in §§ 44b und 60d UrhG können einen urheberrechtlichen Rahmen für den Einsatz von Text und Data Mining Verfahren bieten.[5] Außerhalb dieser Gestattung finden sich in den § 44a ff. UrhG weitere Schranken, die einen gewissen Rahmen – wie beispielsweise das Zitatrecht[6] in § 51 UrhG oder § 60c UrhG – für die Auseinandersetzung mit oder die Nutzung

---

[1] Die Verfasserin Dipl.-Jur. Katharina Erler-Fridgen ist wissenschaftliche Mitarbeiterin am Institut für Recht und Digitalisierung Trier bei Prof. Dr. Benjamin Raue (IRDT, Universität Trier) und arbeitet im interdisziplinären Forschungsprojekt Mining and Modeling Text (MiMoText, Universität Trier).

[2] *Erler-Fridgen*, Kriterien der urheberrechtlichen Schutzfähigkeit von Texten und Sammelwerken, IRDT PAPER*SERIES* Nr. 2.

[3] *Erler-Fridgen*, Datenbanken als Quelle oder Ergebnis von Textanalysen –Datenbankwerkschutz und das Leistungsschutzrecht des Datenbankherstellers, IRDT PAPER*SERIES* Nr. 4.

[4] *Erler-Fridgen*, Die Präsentation von Textteilen als Ergänzung von Textanalysen, IRDT PAPER*SERIES* Nr. 3.

[5] *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6.

[6] *Erler-Fridgen*, Das Zitat und dessen Rahmen für Belege bei Textanalysen, IRDT PAPER*SERIES* Nr. 7.

Institut für Recht und Digitalisierung Trier (IRDT), Direktoren: Prof. Dr. Timo Hebeler, Prof. Dr. Benjamin Raue, Prof. Dr. Peter Reiff, Prof. Dr. Antje von Ungern-Sternberg
Universität Trier, Campus II, Gebäude H, 54286 Trier, Behringstraße 21, 54296 Trier, https://irdt.uni-trier.de

Zitiervorschlag:
Erler-Fridgen, Verfahrensschritte bei dem Einsatz von Text und Data Mining-Verfahren in den Geisteswissenschaften, IRDT PAPER*SERIES* Nr. 5.

von Texten bieten können. Außerhalb dieser Schranken wird es nötig, die Zustimmung des Urhebers zu einer Verwertungshandlung einzuholen. Im Folgenden werden Verfahrensschritte beim Text und Data Mining in den Geisteswissenschaften geschildert und daran anknüpfend urheberrechtlich relevante Handlungen beschrieben.

A. Beschreibung des iterativen Verfahrens

Die Digital Humanities entwickeln und nutzen Verfahren zur Identifikation, Extraktion, Analyse und Vernetzung von fachwissenschaftlich relevanten Informationen aus einschlägigen Datenbeständen. Je nach Projektkontext werden unterschiedliche Text- und Datenquellen gesammelt und aufbereitet, um sodann Fachinformationen zu extrahieren, zu vernetzen und schließlich zu präsentieren. Die urheberrechtliche Perspektive knüpft üblicherweise an menschliche Handlungen an.[7] Daher sind die genannten Handlungen wesentliche Anknüpfungspunkte für die urheberrechtliche Bewertung des Verfahrens beim Text und Data Mining in den Geisteswissenschaften.[8]

Werden bei der Sammlung, Aufbereitung und Auswertung bestehender Texte Kopiervorgänge vorgenommen, so sind diese Verfahrensschritte dazu geeignet, das Vervielfältigungsrecht des Urhebers nach § 16 UrhG zu beeinträchtigen.[9] Dabei werden auch vorübergehende Vervielfältigungen in dem Arbeitsspeicher des Computers als relevante Vervielfältigungen im Sinne des § 16 UrhG angesehen, können jedoch aufgrund der Schranke des § 44a UrhG gestattet sein.[10] Jedoch ist etwa für die Zusammenstellung, Normalisierung und Annotation eines Datenkorpus bei fundierten Textanalysen in der Regel notwendig, die Vervielfältigungen dauerhaft zu speichern, was eine über § 44a UrhG hinausgehende urheberrechtliche Gestattung nötig macht.[11]

Auch die Übernahme von Textteilen kann rechtswidrige Vervielfältigungshandlungen begründen, weil auch Textteile eigenständig urheberrechtlich geschützt sein können.[12] Werden schließlich die Auswertungsergebnisse im Netz präsentiert sowie die Informationen zur

---

[7] *De la Durantaye/Raue*, RuZ 2020, 83, 90, siehe zu Verwertungshandlungen *v. Ungern-Sternberg*, in Schricker/Loewenheim, Urheberrecht, 6. Auflage 2020, § 15 Rn. 10.

[8] So ausführlich auch: *De la Durantaye/Raue*, RuZ 2020, 83, 90.

[9] *De la Durantaye/Raue*, RuZ 2020, 83, 90.

[10] *Schulze*, in Dreier/Schulze, Urheberrechtsgesetz, 7. Aufl. 2022, § 16 Rn. 7; Flüchtige Vervielfältigungshandlungen können jedoch von der urheberrechtlichen Schranke des § 44a UrhG freigestellt sein, zu den Voraussetzungen siehe: *v. Welser*, in Wandtke/Bullinger, Urheberrecht, 5. Aufl. 2019, § 44a Rn. 1.

[11] *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6, II. mit Verweis auf *Raue*, ZUM 2021, 793, 795; siehe auch *Raue*, IIC 2018, 379, 381: vorbereitende Handlungen wie die Digitalisierung, die Normalisierung oder die Annotierung des Korpus verursachen typischerweise eine längere Speicherdauer als sie von Art. 5 Abs. 1 InfoSoc-RL/§ 44a UrhG freigestellt ist.

[12] Zum eigenständigen urheberrechtlichen Schutz von Textteilen *Erler-Fridgen*, Die Präsentation von Textteilen als Ergänzung von Textanalysen, IRDT PAPER*SERIES* Nr. 3.

Nachnutzung durch Dritte bereitgestellt, können – je nach Art und Umfang[13] – Eingriffe in das Recht des Urhebers auf öffentliche Zugänglichmachung nach § 19a UrhG begründet werden.[14]

Die fünf Phasen des wissenschaftlichen Arbeitens[15] können eingesetzt werden, um den Einsatz von Text und Data Mining Verfahren in den Geisteswissenschaften aus urheberrechtlicher Sicht zu beleuchten: Werden solche Textanalyseverfahren genutzt, werden Texte und Daten gesammelt (1), diese aufbereitet (2), Informationen extrahiert (3), die Analyseergebnisse und Text(teile) präsentiert (4) sowie die Forschungsdaten archiviert (5).[16] Im Folgenden sollen diese einzelnen Verfahrensschritte bei der Anwendung von Text und Data Mining Verfahren in den Geisteswissenschaften dargestellt werden. Insbesondere sollen dabei mögliche urheberrechtlich relevante Handlungen identifiziert werden.

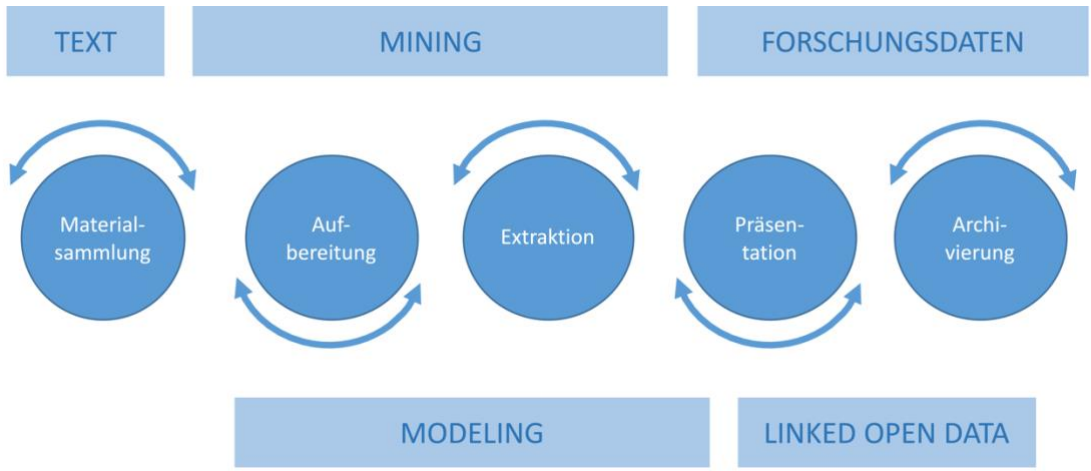

Abbildung 1: „Iterative Verfahrensschritte bei dem Einsatz von Text und Data Mining in den Geisteswissenschaften" von Katharina Erler-Fridgen und Prof. Dr. Benjamin Raue, CC BY-SA 4.0.

Die genannten Verfahrensschritte[17] erfolgen technisch wie konzeptionell in einem iterativen Verfahren, in dem die einzelnen Verfahrensschritte typischerweise ineinandergreifen. Je nach analysierter Textart und Forschungsansatz kann die Reihenfolge der Verfahrensschritte variieren. Beispielsweise kann die (semi-) automatische Informationsextraktion als Teil der Modellierung selbst Voraussetzung zur weiteren Aufbereitung der Textquellen sein (s.u.).

[13] Beispielsweise könnten vollständige Primärtexte oder lediglich Analyseergebnisse veröffentlicht werden. Auch können Textteile als Ergänzung zu extrahierten Informationen präsentiert werden.
[14] *Jotzo*, RuZ 2020, 128, 130.
[15] Ausführlich zu diesen fünf Phasen siehe *de la Durantaye/Raue*, RuZ 2020, 83, 90.
[16] Hierzu grundlegend *de la Durantaye/Raue*, RuZ 2020, 83, 90.
[17] Allgemein zu den fünf Phasen wissenschaftlichen Arbeitens im digitalen Zeitalter siehe: *De la Durantaye/Raue*, RuZ 2020, 83, 90; zu etwas gebündelten Verfahrensschritten: *Specht*, OdW 2018, 285.

B. Die Verfahrensschritte im Einzelnen[18]

I. Sammlung der Text- und Datenquellen

Um Fachinformationen zu gewinnen, die zur weiteren Wissensgenerierung ausgewertet werden sollen, müssen in einem ersten Schritt entsprechende **Text- und Datenquellen** gesammelt werden. Je nach Art der gewünschten Informationen können unterschiedliche Textarten als Quelle genutzt werden. Beispielsweise werden Primärtexte, Fachliteratur sowie bibliographische Datenbestände eingesetzt. Die gesammelten Text- und Datenquellen befinden sich meist in einem unterschiedlichen Aufbereitungszustand und können unterschiedlichen Publikationsquellen entstammen, z.B. Originalausgaben, wissenschaftliche Editionen, Sammelbänden, Bibliographien, Metadatenplattformen, OpenAccess Plattformen etc.

1. Sammlung verschiedener Text- und Datenarten

Für die Sammlung von **Metadaten** literarischer Werke werden beispielsweise bibliographische Verzeichnisse, Werkverzeichnisse eines Urhebers oder bereits vorhandene Datenbanken herangezogen. Auch Metadatendienste der Bibliotheken[19] und Archive können hierfür genutzt werden.

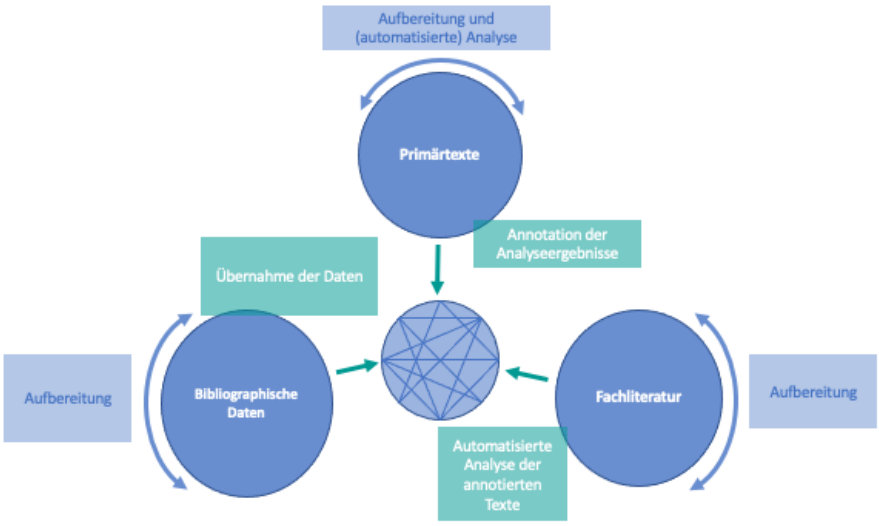

Abbildung 2: Beispiel für die Sammlung, Aufbereitung und Extraktion von unterschiedlichen Textquellen

[18] Dank gilt Prof. Dr. Christof Schöch und Dr. Maria Hinzmann, Projekt MiMoText, Universität Trier, für den Austausch zum Thema.

[19] Bspw. sind über die Metadatendienste der Deutschen Nationalbibliothek alle Titeldaten des Bestands aktualisiert abrufbar als Linked-Data-Service, teilweise mittels Klassifikation der Werke nach Thema siehe https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/LDS/lds_node.html (abgerufen am 12.01.2022); darüber hinaus werden auch Entity-Facts (maschinenlesbare Faktenblätter zur Analyse von Entitäten, bspw. Namensformen) https://www.dnb.de/DE/Professionell/ Metadatendienste/Datenbezug/Entity-Facts/entityFacts_node.html (abgerufen am 12.01.2022) innerhalb des Bibliothekbestands zur Verfügung gestellt.

Zum Beispiel kann eine (semi-)automatisierte Auswertung von Texteigenschaften auf Basis von **Primärtexten** erfolgen, die jedoch in maschinenlesbarer Form vorliegen oder entsprechend aufbereitet werden müssen (siehe Beispiel in Abbildung 2; zur Aufbereitung allgemein unten II). Bereits aufbereitete Primärtexte können in Datenbanken[20] oder auf Open Source Plattformen verfügbar sein.[21] Primärtexte, die noch aufbereitet werden müssen, können der Originalveröffentlichung oder wissenschaftlichen Editionen entnommen werden. Des Weiteren können Analyseergebnisse aus Text und Data Mining-Analysen auch zur weiteren Auswertung annotiert werden (siehe Beispiel in Abbildung 2).

Fachliteratur und fachliterarische Teile wissenschaftlicher Editionen sowie Lexika und sonstige Sekundärliteratur können ebenfalls als Quellen der Fachinformationsgewinnung gesammelt werden. Diese können zum Beispiel nach erfolgter Annotation automatisiert ausgewertet werden (siehe Beispiel in Abbildung 2).

2. Urheberrechtliche Perspektive auf die Sammlung von Texten und Daten

Aus **urheberrechtlicher Perspektive** erzeugt das Sammeln und Speichern einzelner Text- und Datenquellen digitale Kopien und greift damit in das Vervielfältigungsrecht nach § 16 UrhG ein. Sind die Textquellen urheberrechtlich geschützt, so sind sie nach § 64 UrhG für einen Zeitraum von 70 Jahren nach dem Tod des Urhebers vor entsprechenden Verwertungshandlungen geschützt.[22] Texte sind nach § 2 Abs. 2 UrhG urheberrechtlich geschützt, wenn sie eine persönliche geistige Schöpfung in wahrnehmbarer Form darstellen und ausreichende Individualität aufweisen.[23] Darüber hinaus kann das Sammeln von Werken oder Daten aus Sammelbänden, etwa einer bibliographischen Sammlung, in den urheberrechtlichen Schutz des Sammelwerks nach § 4 UrhG eingreifen.[24]

Frei nutzbar sind grundsätzlich gemeinfreie Werke nach § 64 UrhG und ggf. verwaiste Werke nach § 61 UrhG.[25] Urheberrechtlich geschützte Werke können mit Zustimmung des Urheberrechtsinhabers beispielsweise über Creative-Commons-Lizenzen genutzt werden, wobei

---

[20] Zur Entnahme einzelner Elemente aus vorhandenen Datenbanken siehe *Erler-Fridgen*, Datenbanken als Quelle oder Ergebnis von Textanalysen – Datenbankwerkschutz und das Leistungsschutzrecht des Datenbankherstellers, IRDT PAPER*SERIES* Nr. 4.

[21] Dafür kommen etwa Sammlungen gemeinfreier Primärtexte der Staatsbibliotheken in Betracht, zum Beispiel die der französischen Nationalbibliothek Gallica: https://gallica.bnf.fr/accueil/de/content/accueil-de?mode=desktop (abgerufen am 12.01.2022), auch Repositorien, die auf Open-Source Plattformen wie github (s.o.) veröffentlicht wurden, könnten dafür eingesetzt werden.

[22] *Erler-Fridgen*, Kriterien der urheberrechtlichen Schutzfähigkeit von Texten und Sammelwerken, IRDT PAPER*SERIES* Nr. 2, I. 1.

[23] Ausführlich hierzu: *Erler-Fridgen*, Kriterien der urheberrechtlichen Schutzfähigkeit von Texten und Sammelwerken, IRDT PAPER*SERIES* Nr. 2, I. 1.

[24] Zum urheberrechtlichen Schutz von Sammelwerken siehe *Erler-Fridgen*, Kriterien der urheberrechtlichen Schutzfähigkeit von Texten und Sammelwerken, IRDT PAPER*SERIES* Nr. 2.

[25] *Hagemeier*, in Ahlberg/Götting/Lauber-Rönsberg, BeckOK Urheberrecht, 33. Edition 2022, § 60d Rn. 14; siehe zur Gemeinfreiheit auch *Erler-Fridgen*, Kriterien der urheberrechtlichen Schutzfähigkeit von Texten und Sammelwerken, IRDT PAPER*SERIES* Nr. 2, II. 1.

jedoch deren Voraussetzungen beachtet werden müssen. Die Übersetzung an sich gemeinfreier Werke kann allerdings als Bearbeitung nach § 3 UrhG geschützt sein.[26]

Außerdem kann die Sammlung o.g. verschiedener Texte und Daten aus Datenbanken einerseits den urheberrechtlichen Schutz des Datenbankwerks nach § 4 Abs. 2 UrhG verletzen.[27] Andererseits kann durch eine Entnahme auch in das *sui generis* Recht des Datenbankherstellers nach § 87a ff. UrhG eingegriffen werden, wenn wesentliche Teile einer durch wesentliche Investition geschaffenen Datenbank genutzt werden.[28]

Die Nutzung wissenschaftlicher Editionen als Quelle für Primärtexte oder fachliterarischer Ergänzungsteile kann das Leistungsschutzrecht der wissenschaftlichen Ausgabe nach § 70 UrhG verletzen.[29] Solche wissenschaftliche Ausgaben sind nach § 70 UrhG für 25 Jahre nach ihrem Erscheinen leistungsschutzrechtlich geschützt.

Die Vervielfältigungen können bei rechtmäßigem Zugang zu den Werken auf Grundlage der Text und Data Mining-Schranken (§ 44b, 60d UrhG) zulässig sein.[30] § 44b Abs. 2 UrhG stellt es frei, bei rechtmäßigem Zugang zu den Werken, diese für die weitere Aufbereitung und Extraktion zusammenzustellen.[31] Ebenso umfasst § 60d UrhG bei rechtmäßigem Zugang das Recht, die Quellen zu speichern, um sie sodann aufzubereiten.[32]

II. Aufbereitung

Die verschiedenen Text- und Datenquellen müssen auf Basis einer **Modellierung** aufbereitet werden, um Texte mit informationstechnischen Mitteln untersuchen zu können. Als Grundlage der Aufbereitung erfolgen eine konzeptionelle Modellierung sowie eine Modellierung mit Blick auf die technische Umsetzung der eingesetzten Verfahren.[33]

1. Modellierung

Ziel der **Modellierung** in den Digital Humanities ist es, relevante Elemente und Merkmale der Texte zu identifizieren. Diese sind dann die Basis für Mining-Analyseverfahren. Hierfür werden Modelle eingesetzt, die die ausgewählten Elemente der Texte maschinenlesbar repräsentieren.[34] Die

---

[26] *De la Durantaye/Raue*, RuZ 2020, 83, 91.

[27] *Erler-Fridgen*, Datenbanken als Quelle oder Ergebnis von Textanalysen – Datenbankwerkschutz und das Leistungsschutzrecht des Datenbankherstellers, IRDT PAPER*SERIES* Nr. 4.

[28] Näher zum sui generis Schutz des Datenbankherstellers siehe *Erler-Fridgen*, Datenbanken als Quelle oder Ergebnis von Textanalysen – Datenbankwerkschutz und das Leistungsschutzrecht des Datenbankherstellers, IRDT PAPER*SERIES* Nr. 4; siehe auch *Spindler*, GRUR 2016, 1112, 1113 f.

[29] *Erler-Fridgen*, Die Nutzung wissenschaftlicher Ausgaben für Textanalysen, IRDT PAPER*SERIES* Nr. 1.

[30] *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6; *Hagemeier*, in Ahlberg/Götting/Lauber-Rönsberg, BeckOK Urheberrecht, 33. Edition 2022, § 60d Rn. 15.

[31] *Raue*, ZUM 2021, 793, 795.

[32] *Raue*, ZUM 2021, 793, 797.

[33] Mit Blick auf die technische Umsetzung wird teilweise auch von logischer Modellierung gesprochen, siehe *Jannidis*, Grundlagen der Datenmodellierung, in Jannidis/Kohle/Rehbein (Hrsg.), Digital Humanities, S. 103.

[34] *Jannidis*, Grundlagen der Datenmodellierung, in Jannidis/Kohle/Rehbein (Hrsg.), Digital Humanities, S. 100 ff.

Auswahl der repräsentierten Elemente erfolgt konzeptionell und wird je nach Verwendungszweck für spätere Textanalysen getroffen.[35] Dabei liegt ein wesentlicher Forschungsschwerpunkt der Modellierung darin, dass sich verschiedene geisteswissenschaftliche Disziplinen auf eine für das jeweilige Forschungsvorhaben adäquate Modellierung verständigen.[36] Zugleich sollen Konzepte und Standards der Modellierung gewährleisten, dass Daten auch über den Projektkontext hinaus nutzbar sind. Es soll möglich sein, dass auf der entstandenen Datenbasis mit anderen Forschern zusammengearbeitet werden kann.[37] Außerdem ist oftmals vorgesehen, dass die geplante Präsentation der Daten in die Konzeption der Modellierung miteinfließt, insbesondere dadurch, dass Modelle und technische Verknüpfungsstrategien standardisiert werden.[38]

Auf Basis der konzeptionellen wie technischen Modellierung erfolgt sodann die **maschinenlesbare Aufbereitung** mit Hilfe verschiedener Verfahren[39], die Auswahl der passenden Formate[40] sowie die Annotation der Texte.

2. Annotationen

Annotation meint das Setzen von ergänzenden Hinweisen, also die Kommentierung einzelner Token im Text als sogenannter Beitext.[41] Je nach Einsatz stellen **Annotationen** einerseits einen wesentlichen wissenschaftlichen Zwischenschritt zwischen Aufbereitung und Informationsextraktion dar. Andererseits können mit ihrer Hilfe auch Ergebnisse der

---

[35] Zur Auswahl und Strukturierung der Daten als „ersten hermeneutischen Akt" siehe *Leif Scheuermann*, Die Abgrenzung der digitalen Geisteswissenschaften, in Digital Classics Online 2 (2016), 1, S. 61, https://journals.ub.uni-heidelberg.de/index.php/dco/article/view/22746/21865 (abgerufen am 12.01.2022) sowie auch: *Deck*, Digital Humanities – Eine Herausforderung an die Informatik und an die Geisteswissenschaften, in: Huber/Krämer (Hrsg.), Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden, http://zfdg.de/sb003_002 (abgerufen am 12.01.2022).

[36] *Thaller*, Digital Humanities als Wissenschaft, in Jannidis/Kohle/Rehbein (Hrsg.), Digital Humanities, S. 16, *McCarty*, Modeling: A Study in Words and Meanings, in Ray Siemens, John Unsworth, and Susan Schreibman, A Companion to Digital Humanities, http://www.digitalhumanities.org/companion/ (abgerufen am 12.01.2022).

[37] *Scheuermann*, Die Abgrenzung der digitalen Geisteswissenschaften, in Digital Classics Online 2 (2016), 1, S. 61, https://journals.ub.uni-heidelberg.de/index.php/dco/article/view/22746/21865 (abgerufen am 12.01.2022).

[38] Zu den dahingehenden Funktionen der Datenmodelle: *Jannidis*, Grundlagen der Datenmodellierung, in Jannidis/Kohle/Rehbein (Hrsg.), Digital Humanities, S. 100.

[39] Optical Character Recognition (OCR): Verfahren zur Textdigitalisierung, in dem der Prozess der Texterfassung automatisiert vorgenommen wird, siehe http://digitalhumanities.berkeley.edu/resources/digitization-workflows-scanning-ocr-and-audio-transcription (abgerufen am 12.01.2022) und Double Keying Verfahren: Double Keying beschreibt ein Verfahren zur Textdigitalisierung, in dem der Prozess der Texterfassung manuell vorgenommen wird, siehe http://www.dhmuseum.uni-trier.de/node/49 (abgerufen am 12.01.2022).

[40] Beispielsweise werden Texte im Format XML (Extensible Markup Language, eine Auszeichnungssprache, die zur Annotation eingesetzt werden kann und als Datenaustauschformat eine einheitliche Repräsentation in unterschiedlichen Applikationskontexten ermöglicht, siehe *Rhem*, 2.5 Texttechnologische Grundlagen, in K.-U. Carstensen, Ch. Ebert, C. Ebert, S. Jekat, R. Klabunde, H. Langer (Hrsg.), Computerlinguistik und Sprachtechnologie, 3. Aufl., S. 159.) oder im Format RDF (Resource Description Framework, ist ein Standard zur Annotation von Metadaten in der Form von Triplets (Subjekt, Prädikat, Objekt), *Rhem*, 2.5 Texttechnologische Grundlagen, in K.-U. Carstensen, Ch. Ebert, C. Ebert, S. Jekat, R. Klabunde, H. Langer (Hrsg.), Computerlinguistik und Sprachtechnologie, 3. Aufl., S. 166) erschlossen.

[41] *Rapp*, Manuelle und automatische Annotation, in Jannidis/Kohle/Rehbein (Hrsg.), Digital Humanities, S. 253 f.

Informationsextraktion aufgenommen werden. Zum Teil werden Annotationen manuell vorgenommen, etwa um technische Merkmale des Digitalisats sowie linguistische Merkmale des Textes hervorzuheben. Beispielsweise können dabei die Haltung des Aussagegebenden zum Satz, Quellenangaben sowie Aussagen zur Beziehung aus Kookkurrenz[42] manuell annotiert werden. Weiterhin werden automatisiert extrahierte Informationen durch Annotationen als Kommentierung am Text ergänzt. Hier zeigt sich der iterative Prozess der Verfahrensschritte, Aufbereitung und Informationsextraktion, an deren Schnittstellen Annotationen genutzt werden.

3. Urheberrechtliche Perspektive auf die Aufbereitung

Bei der Aufbereitung der gesammelten Text- und Datenbestände können urheberrechtlich relevante Handlungen vorgenommen werden. Werden nämlich in diesem Verfahrensschritt Kopien erzeugt oder die Dateien im Arbeitsspeicher zwischengespeichert, so kann dies in das Vervielfältigungsrecht nach § 16 UrhG eingreifen. Beispielsweise können das Scannen der Texte,[43] das Digitalisieren[44] der Texte, etwa mittels Optical Charakter Recognition (OCR)[45], aber auch das Speichern[46] nach erfolgter Annotation der Texte oder die Umwandlung der Dateiformate relevante Vervielfältigungshandlungen nach § 16 UrhG sein. Solche Vervielfältigungen zur Aufbereitung von Texten sind jedoch freigestellt, wenn die Voraussetzungen der Text und Data Mining-Schranken (§ 44b/§ 60d UrhG) vorliegen.[47] Daneben stellt § 44a UrhG lediglich vorrübergehende Vervielfältigungen frei, die der technischen Übermittlung dienen und keine eigene wirtschaftliche Bedeutung aufweisen.[48] Insbesondere bedeutet vorübergehend in diesem Kontext, dass die Daten nach einer nicht ins Gewicht fallenden Zeit automatisch wieder gelöscht werden müssen.[49]

a. Maschinenlesbare Aufbereitung

Wie geschildert, reicht jedoch die reine Vervielfältigung in Form von Kopien oftmals nicht aus, sondern es kann notwendig sein, die Texte etwa in eine maschinenlesbare Form umzuwandeln.[50] Würde das eine Bearbeitung darstellen, dann wäre diese zwar ohne Zustimmung der

---

[42] Als Kookkurrenz wird das gemeinsame Vorkommen zweier oder mehrerer Wörter in einem Kontext von fest definierter Größe bezeichnet, *Kunze/Lemnitzer*, Computerlexikographie. Eine Einführung, S. 391f.; Näheres zur Kookkurenzanalyse: *Engelberg*, Kookkurrenzanalyse, Linguistische Methodenlehre, FS 2009, Uni Mannheim, https://www1.ids-mannheim.de/fileadmin/lexik/lehre/engelberg/Webseite_LingMeth /Skript_05.pdf. (abgerufen am 12.01.2022).

[43] BGH GRUR 2002, 246, 247 – Scanner.

[44] Zur Erzeugung eines Datensatzes mittels eines Digitalisierungsgeräts: KG ZUM 2002, 828, 830 – Pressespiegelversand.

[45] Verfahren zur Textdigitalisierung, in dem der Prozess der Texterfassung automatisiert vorgenommen wird, siehe http://digitalhumanities.berkeley.edu/resources/digitization-workflows-scanning-ocr-and-audio-transcription (abgerufen am 10.03.2020).

[46] Landgericht München I ZUM-RD 2003, 607, 610 – Vervielfältigung von Musiktiteln.

[47] *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6.

[48] *Dreier*, in Dreier/Schulze, Urheberrechtsgesetz, 7. Aufl. 2022, § 44a Rn. 4 ff.

[49] EuGH C-5/08, ECLI:EU:C:2009:465, GRUR Int 2010, 35 Rn. 64 – Infopaq; *Dreier*, in Dreier/Schulze, Urheberrechtsgesetz, 7. Aufl. 2022, § 44a Rn. 4.

[50] *Spindler*, GRUR 2016, 1112, 1113.

rechteinhabenden Person möglich, jedoch wäre nach § 23 UrhG die Verwertung, also auch die Vervielfältigung, nicht ohne eine solche Zustimmung zulässig.[51] Das Gesetz besagt jedoch in § 23 Abs. 3 UrhG, dass „ausschließlich technisch bedingte Änderungen eines Werkes nach § 44b Abs. 2, 60d Abs. 1, § 60e Abs. 1 sowie § 60f Abs. 2 UrhG" als Bearbeitung nach § 23 UrhG nicht erfasst werden.[52] Damit stellt § 23 Abs. 3 UrhG klar, dass die Digitalisierung[53], die Umwandlung in ein maschinenlesbares Format sowie die rein technisch bedingte Anreicherung mit Metadaten wohl keine Bearbeitung im Sinne des § 23 UrhG darstellen.[54] Hierfür spricht auch, dass insbesondere bei der Digitalisierung das geistige Wesen des Werkes unverändert bleibt.[55] Wie diese Norm im Verhältnis zum unionsrechtlichen Vervielfältigungsbegriff steht, wird in der rechtswissenschaftlichen Literatur diskutiert und ist nicht abschließend geklärt.[56]

b. Rechte der Forschenden durch Aufbereitung

Ob durch die Aufbereitung von Texten eigene Rechte für die Forschenden entstehen, ist fraglich und wird in vielen Forschungsgruppen eine wichtige Rolle spielen.[57] Gerade Anreicherungen mit Metainformationen wie Annotationen werden regelmäßig in Forschungsprojekten unter hohem Aufwand erstellt. Für den urheberrechtlichen Schutz von Annotationen als Bearbeitung (§ 3 UrhG) wäre erforderlich, dass die Aufbereitung die Schwelle hinreichender Individualität überschreitet.[58] Für Ergänzungen an einem vorhandenen Text gilt: Je mehr sich der Gestaltungsspielraum durch Ergänzungen am Text erweitert, desto eher könnte Individualität vorliegen.[59] Ob Annotationen als Ergänzungen oder Abänderungen des Ausgangstextes und bei hinreichender Individualität als schutzfähige Bearbeitungen gewertet werden können, ist zweifelhaft. Teilweise wird vertreten, dass linguistische Annotationen bei erforderlicher Schöpfungshöhe Bearbeitungen im Sinne von § 3 UrhG darstellen können.[60] Hierbei müsse jedoch zwischen vorbereitenden Annotationsschemata, entsprechender Standardsetzung und der Durchführung der Annotation am Text unterschieden werden: es sei denkbar, dass Annotationsschemata Schöpfungshöhe erreichen, die Durchführung am Text als solche jedoch nicht.[61] Anderer Ansicht nach liegt bei der Anreicherung mit

---

[51] *Spindler*, GRUR 2016, 1112, 1113.
[52] *De la Durantaye/Raue*, RuZ 2020, 83, 91 f.
[53] *Loewenheim,* in Schricker/Loewenheim, Urheberrecht, 6. Auflage 2020, § 23 Rn. 7; *Ernst*, in Hoeren/Sieber/Holznagel, Multimedia-Recht, 7.1 Rn. 51.
[54] *Specht*, OdW 2018, 285.
[55] Siehe zur Digitalisierung und Bearbeitung allgemein *Loewenheim,* in Schricker/Loewenheim, Urheberrecht, 6. Auflage 2020, § 23 Rn. 7.
[56] *De la Durantaye/Raue*, RuZ 2020, 83, 92; ausführlich *Raue*, AfP 2022, 1 ff.; zur Diskussion *Ohly,* GRUR 2017, 964, 967 mwN: Der Ansicht nach, der die ständige Rechtsprechung des BGH entspreche, wäre die Bearbeitung ein Unterfall der unionsrechtlichen Vervielfältigung.
[57] *De la Durantaye/Raue*, RuZ 2020, 83, 92.
[58] *Schulze*, in Dreier/Schulze, Urheberrechtsgesetz, 7. Aufl. 2022, § 3 Rn. 11.
[59] *Schulze*, in Dreier/Schulze, Urheberrechtsgesetz, 7. Aufl. 2022, § 3 Rn. 17.
[60] *Lehmberg, Chiarcos, Rhem, Witt*, Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten, in Rehm/Witt/Lemnitzer, Datenstrukturen für linguistische Ressourcen und ihre Anwendungen, S. 93, 98.
[61] *Lehmberg, Chiarcos, Rhem, Witt*, Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten, in Rehm/Witt/Lemnitzer, Datenstrukturen für linguistische Ressourcen und ihre Anwendungen, S. 93, 98.

Metainformationen keine Bearbeitung vor.[62] Denn die angefügten Metainformationen würden nicht das eigentliche Werk berühren, sondern es nur anreichern und schematisieren.[63] Hierin liege gerade keine Veränderung des ursprünglichen Werkes, die Informationen charakterisieren und ordnen das Werk lediglich ein.[64] Es liegt keine höchstrichterliche Rechtsprechung zur Beurteilung von linguistischen Annotationen als Bearbeitungen bzw. deren Schöpfungshöhe vor.[65] Festzuhalten ist: eine Bearbeitung erfordert eine Veränderung der schutzfähigen Merkmale des ursprünglichen Werks, während die Übernahme ohne Veränderung eine Vervielfältung darstellt.[66] In Abgrenzung dazu legt § 23 Abs. 3 UrhG jedenfalls die Annahme nahe, dass eine ausschließlich technisch bedingte Anreichung mit Metadaten ebenso wie die Digitalisierung bereits keine Bearbeitung im Sinne des § 23 Abs. 1 UrhG darstellt.[67]

III. Informationsextraktion

Die Informationsextraktion kann mittels verschiedener Analysemethoden durchgeführt werden. Zunächst geht es beim Text und Data Mining darum, verborgene Zusammenhänge und Strukturen im Text aufzufinden. Quantitative **Verfahren** der Informationsextraktion sind beispielsweise statistische Analysen, das maschinelle Lernen, Netzwerkanalysen sowie das Topic Modeling[68] und die Sentiment Analyse[69].[70] Je nach Art der Textquelle können im konkreten Projektkontext unterschiedliche Analyseverfahren eingesetzt werden. Auf Grundlage der Ergebnisse – etwa nach der Topic Modeling Analyse – können die Texte dann wiederum annotiert werden. Annotierte Texte können umgekehrt aber auch Trainingsdaten für weiterführende Analysen und damit die Grundlage für zusätzliche Informationsextraktion bieten. Ergänzend kann die Auffindbarkeit bereits enthaltener Informationen mittels Information Retrieval sichergestellt werden.[71]

Grundsätzlich ist die Extraktion von Informationen aus urheberrechtlich geschützten Quellen als solche nicht vom Zuweisungsgehalt des Urheberrechts erfasst.[72] Grundsätzlich gilt dies auch dann, wenn die Extraktion mit technischen Mitteln wie etwa Algorithmen erfolgt.[73] Erfordert die Extraktion von Informationen es jedoch, dass Kopien erzeugt oder die Dateien in den

---

[62] *Spindler*, GRUR 2016, 1112, 1114.

[63] *Spindler*, GRUR 2016, 1112, 1114: Dies entspreche in der analogen Welt am ehesten einer Exzerption von Teilen eines Werkes mit Hilfe bestimmter Schlagwörter oder Indizes.

[64] *Spindler*, Text und Data Mining – urheber- und datenschutzrechtliche Fragen, GRUR 2016, 1112, 1114.

[65] *Lehmberg, Chiarcos, Rhem, Witt*, Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten, in Rehm/Witt/Lemnitzer, Datenstrukturen für linguistische Ressourcen und ihre Anwendungen, S. 93, 98.

[66] *Loewenheim,* in Schricker/Loewenheim, Urheberrecht, 6. Auflage 2020, § 23 Rn. 6.

[67] *Specht*, OdW 2018, 285.

[68] *Blei*, Probabilistic Topic Models, http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf (abgerufen am 12.01.2022).

[69] *Ignatow/Mihalcea*, Text Mining, S. 148 ff.

[70] Ausführlich zu den unterschiedlichen Verfahrensmethoden siehe *Schöch*, Quantitative Analyse, in Jannidis/Kohle/Rehbein (Hrsg.), Digital Humanities, S. 279 ff.

[71] *Klinke*, Information Retrieval, in Jannidis/Kohle/Rehbein (Hrsg.), Digital Humanities, S. 268 f.

[72] *Raue,* ZUM 2019, 684, 686.

[73] *Raue,* GRUR 2017, 11, 13; *De la Durantaye/Raue*, RuZ 2020, 83, 92.

Arbeitsspeicher geladen werden, was in der Regel der Fall sein wird, so kann aus **urheberrechtlicher Perspektive** auch dieser Verfahrensschritt zu Vervielfältigungshandlungen nach § 16 UrhG führen.[74] Für das Text und Data Mining vorgenommenen Vervielfältigungshandlungen fallen unter die Text und Data Mining-Schranken nach § 44b UrhG oder § 60d UrhG und sind unter deren Voraussetzungen freigestellt.[75] Voraussetzung ist jedoch insbesondere der rechtmäßige Zugang zu den genutzten Werken.[76] Hiernach zulässig sind Vervielfältigungen zum Zweck des Text und Data Minings, sodass neben der automatisierten Analyse, die Quellen auch zuvor gespeichert, normalisiert, annotiert oder auf sonstige Weise bearbeitet (§ 23 Abs. 3 UrhG) werden können (zur Aufbereitung siehe III.).[77] Die Schranke der vorübergehenden Vervielfältigungen nach § 44a UrhG setzt hingegen voraus, dass die Kopien zeitlich beschränkt sind, der technischen Übermittlung dienen und keine eigene wirtschaftliche Bedeutung aufweisen.[78]

IV. Präsentation

Die Präsentation der extrahierten Informationen umfasst mehrere Grundentscheidungen. In welchem Umfang sollen die Daten veröffentlicht werden? Wie frei sollen die Daten verfügbar sein? Sind personenbezogene Daten mitenthalten?[79] Wie soll die Benutzeroberfläche gestaltet sein? In welchem Format sollen die Daten veröffentlicht werden?

Dabei spielt die **Nachnutzbarkeit** vorhandener Forschungsergebnisse bzw. die Kombinationsmöglichkeit mit vorhandenen Datenbeständen eine wesentliche Rolle. Insbesondere Anbindungsmöglichkeiten der gewonnenen Informationen an vorhandene externe Netzwerke können die Nachnutzbarkeit der Informationen erhöhen. Beispielsweise können extrahierte Daten als Linked Open Data verfügbar gemacht werden, wodurch die Vernetzung mit weiteren Informationen möglich ist.[80] Nachnutzungsoptionen frei verfügbarer Datenbestände durch Dritte müssen dabei mitgedacht werden: Sollen die Datenbestände durch andere Nutzer ergänzbar sein?[81] Werden im Projekt beispielsweise Tools auf den Datensätzen trainiert,[82] dann steht auch die

---

[74] *Schulze*, in Dreier/Schulze, Urheberrechtsgesetz, 7. Aufl. 2022, § 16 Rn. 7.

[75] Zur Zulässigkeit von Vervielfältigungen bei Text und Data Mining siehe *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6; siehe zur Freistellung *de la Durantaye/Raue*, RuZ 2020, 83, 92.

[76] *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6.

[77] *Raue*, ZUM 2021, 793, 797 und 798.

[78] *Dreier*, in Dreier/Schulze, Urheberrechtsgesetz, 7. Aufl. 2022, § 44a Rn. 4 ff.: in zeitlicher Hinsicht müssen die Daten nach einer nicht ins Gewicht fallenden Zeit automatisch gelöscht werden.

[79] Auf mögliche datenschutzrechtliche Implikationen wird in der vorliegenden Handreichung nicht eingegangen.

[80] Linked Open Data bedeutet, dass auf einer Ontologie basierte Datenbanken veröffentlicht werden und zur Vernetzung die Technologie des Seminatik Web genutzt wird, *Rehbein*, Ontologien in den Digital Humanities, in Jannidies/Kohle/Rehbein (Hrsg.), Digital Humanities, S. 173.

[81] Beispielsweise könnte die Einbindung in ein bestehendes Wissensnetzwerk wie das Wikidata-Netzwerk erwogen werden, https://www.wikidata.org/wiki/Wikidata:Introduction (abgerufen am 12.01.2022).

[82] Beispielsweise kann für die Verbesserung der maschinenlesbaren Aufbereitung mittels Optical Recognition Verfahren ein Tool auf den Texten trainiert werden.

Präsentation eines solchen Tools und die Möglichkeit in Frage, dieses in anderen Kontexten einzusetzen.

Werden die Ergebnisse der Analyse präsentiert und veröffentlicht, so kann aus **urheberrechtlicher Perspektive** das Recht auf öffentliche Zugänglichmachung nach § 19a UrhG beeinträchtigt werden. Insbesondere wenn die Volltexte oder längere Textteile mitveröffentlicht werden, steht eine solche Verletzung in Rede. Eine Präsentation der zugrundliegenden Texte kann für die Forschenden interessant sein, da diese als Nachweise[83] der gewonnenen Erkenntnisse sowie zur Nachnutzung dienen können.[84] Der hierfür vorausgesetzte urheberrechtliche Schutz entfällt jedoch bei gemeinfreien Texten, wie beispielsweise Romanen aus der zweiten Hälfte des 18. Jahrhunderts, wegen Zeitablaufs (§ 64 UrhG).[85] Werden Bestandteile wissenschaftlicher Ausgaben mitveröffentlicht, kann hierdurch das Leistungsschutzrecht wissenschaftlicher Ausgaben nach § 70 UrhG verletzt werden.[86] Werden Textteile, wie einzelne Kapitel, Seiten oder auch Sätze, veröffentlicht, kann das Urheberrechte verletzen, sofern diese Werkteile urheberrechtlich nach § 2 UrhG geschützt werden.[87] Die Text und Data Mining Schranke des § 60d Abs. 4 UrhG ermöglicht die öffentliche Zugänglichmachung[88] von Vervielfältigungen für einen abgegrenzten Kreis von Personen zum Zweck deren gemeinsame wissenschaftliche Forschung.[89] Außerdem stellt sie dies auch einzelnen Dritten zur Überprüfung der Qualität der wissenschaftlichen Forschung frei.[90] Hier ist jedoch keine allgemein zugängliche Präsentation, sondern nur eine an einen abgegrenzten Personenkreis bzw. an einzelne Dritte vorgenommene Präsentation möglich.

Außerdem ist es möglich, dass an einer Sammlung von Informationen das *sui generis*-Leistungsschutzrecht des Datenbankherstellers nach § 87a ff. UrhG – u.a. bei wesentlicher Investition – entsteht.[91] Liegt eine schöpferische Leistung in Auswahl oder Anordnung der

---

[83] Zum Zitatrecht und Nachweisen von Textteilen bei Textanalysen siehe *Erler-Fridgen*, Das Zitat und dessen Rahmen für Belege bei Textanalysen, IRDT PAPER*SERIES* Nr. 7.

[84] *Spindler*, GRUR 2016, 1112, 1113.

[85] Zum urheberrechtlichen Schutz siehe *Erler-Fridgen*, Kriterien der urheberrechtlichen Schutzfähigkeit von Texten und Sammelwerken, IRDT PAPER*SERIES* Nr. 2.

[86] Zum Leistungsschutzrecht wissenschaftlicher Ausgaben siehe Erler-Fridgen, Die Nutzung wissenschaftlicher Ausgaben für Textanalysen, IRDT PAPERSERIES Nr. 1.

[87] Zum urheberrechtlichen Schutz von Textteilen siehe *Erler-Fridgen*, Die Präsentation von Textteilen als Ergänzung von Textanalysen, IRDT PAPER*SERIES* Nr. 3.

[88] In der Regel ist bei der Zugänglichmachung für einen abgegrenzten Personenkreis ohnehin kein öffentliches Zugänglichmachen iSv. § 19a UrhG gegeben: *Raue*, ZUM 2021, 793, 799; zum Begriff der „öffentlichen Zugänglichmachung" auch *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6, IV.

[89] Zu § 60d Abs. 4 UrhG und dem Begriff der „öffentlichen Zugänglichmachung" siehe *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6, IV.

[90] Siehe *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6.

[91] Siehe hierzu *Erler-Fridgen*, Datenbanken als Quelle oder Ergebnis von Textanalysen – Datenbankwerkschutz und das Leistungsschutzrecht des Datenbankherstellers, IRDT PAPER*SERIES* Nr. 4.

Datenbankelemente vor, so kommt auch ein Schutz der präsentierten Sammlung als Datenbankwerk nach § 4 Abs. 2 UrhG in Betracht.[92]

V. Archivierung

Die Archivierung der extrahierten, analysierten und vernetzten Forschungsergebnisse und Daten ist Voraussetzung für die Überprüfbarkeit und Nachnutzbarkeit der erzielten **Forschungsdaten**. Mit dem Ziel der Langzeitarchivierung soll die langfristige Nachnutzbarkeit von Daten trotz technischen Wandels ermöglicht werden.[93] Auch hier sind viele grundsätzliche Entscheidungen zu treffen: In welchem Umfang und auf Basis welchen Projektstandes wird archiviert? Werden personenbezogene Forschungsdaten gespeichert?[94] Welche Metadaten, beispielsweise zu angewandten Forschungsmethoden, Soft- und Hardware etc., werden mitarchiviert? Hierbei müssen auch die passenden Formate der Metadatensätze und Archivdatensätze ausgewählt werden.[95]

Sofern bei der Archivierung der erzielten Forschungsdaten auch Kopien erzeugt werden und (Teile der) Ausgangstexte davon betroffen sind, kann auch an dieser Stelle aus **urheberrechtlicher Perspektive** in das Vervielfältigungsrecht nach § 16 UrhG eingegriffen werden. Die Archivierung kann beispielsweise auf einem projekt-eigenen Server oder mittels Cloud-Repositorien (geschlossen oder frei verfügbar)[96] vorgesehen sein. Werden die Forschungsdaten wie auch Volltexte oder Textteile frei verfügbar archiviert, so kann auch hierdurch in das Recht auf öffentliche Zugänglichmachung nach § 19a UrhG eingegriffen werden. Auch die Zugänglichmachung der Vervielfältigungen für Text und Data Mining an einen abgeschlossenen Kreis – wie sie die Text und Data Mining-Schranke nach § 60d Abs. 4 UrhG zulässt[97] – ist nach Abschluss der gemeinsamen wissenschaftlichen Forschung bzw. Überprüfung der Qualität der wissenschaftlichen Forschung zu beenden.[98] Nach der alten Rechtslage in § 60d UrhG alte Fassung (a.F.) musste Ursprungsmaterial und Korpus nach Abschluss der Forschungsgarbeiten gelöscht werden.[99] Nunmehr ist jedoch eine Archivierung der Vervielfältigungen bei der Forschungseinrichtung selbst nach § 60d Abs. 5 UrhG möglich. Es ist hiernach zulässig, entsprechende Vervielfältigungen mit angemessenen Sicherheitsvorkehrungen aufzubewahren, solange sie für die Zwecke der wissenschaftlichen

---

[92] Siehe *Erler-Fridgen*, Datenbanken als Quelle oder Ergebnis von Textanalysen – Datenbankwerkschutz und das Leistungsschutzrecht des Datenbankherstellers, IRDT PAPER*SERIES* Nr. 4.

[93] https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/langzeitarchivierung/ (abgerufen am 12.01.2022).

[94] Auf mögliche datenschutzrechtliche Implikationen wird in der vorliegenden Handreichung nicht eingegangen.

[95] *Schumann*, Einführung in die digitale Langzeitarchivierung, S. 44 https://www.ssoar.info/ssoar/bitstream/handle/document/45740/ssoar-2012-schumann-Einfuhrung_in_die_digitale_Langzeitarchivierung.pdf (abgerufen am 12.01.2022).

[96] Cloud-Repositorien sind beispielsweise die OpenAccess Cloud Zenodo (https://about.zenodo.org/, abgerufen am 12.01.2022) oder die geschlossen oder frei zugänglich nutzbare Infrastruktur der github-Plattform.

[97] Im Regelfall ist im Zugänglichmachen an einen abgegrenzten Personenkreis jedoch wohl kein öffentliches Zugänglichmachen nach § 19a UrhG zu sehen: *Raue*, ZUM 2021, 793, 799.

[98] Zur Text und Data Mining Schranke siehe *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6.

[99] *Raue*, ZUM 2021, 793, 799.

Forschung oder zur Überprüfung wissenschaftlicher Erkenntnisse erforderlich sind.[100] Das bedeutet, dass Datenkorpora für Anschlussforschungen genutzt werden können, jedoch als Voraussetzung angemessene Sicherheitsvorkehrungen gegen Missbrauch getroffen werden müssen.[101] Aus diesem Grund wird in der rechtswissenschaftlichen Literatur empfohlen, die Daten in Forschungsdatenrepositorien zu speichern, die die Sicherheitsstandards zur Verfügung stellen können.[102] Entsprechend den Grundsätzen guter wissenschaftlicher Praxis ist die Aufbewahrung zur Überprüfung wissenschaftlicher Erkenntnisse jedenfalls zehn Jahre lang erforderlich.[103] Dass eine längere Aufbewahrungsdauer erforderlich ist, müssen die Forschenden beispielsweise unter Verweis auf Anschlussforschung ausreichend plausibel machen.[104]

VI. Ergebnis

Die iterativen Verfahrensschritte – Sammlung, Aufbereitung, Extraktion, Präsentation und Archivierung – greifen beim Text und Data Mining ineinander. Insbesondere die Modellierung setzt konzeptionell das Mitdenken der übrigen Verfahrensschritte voraus. Auch die Annotationen können die Schritte der Aufbereitung und Extraktion miteinander verschränken.

In jedem Verfahrensschritt sind urheberrechtlich relevante Handlungen – Vervielfältigungen nach § 16 UrhG oder das öffentliche Zugänglichmachen nach § 19a UrhG – zu identifizieren. Vervielfältigungen, die für das Text und Data Mining vorgenommen werden, sind nach §§ 44b/60d UrhG unter deren Voraussetzungen zulässig. Die Text und Data Mining-Schranken ermöglichen jedoch nicht, Ausgangstexte allgemein zu präsentieren. Nach § 60d Abs. 4 UrhG ist lediglich die öffentliche Zugänglichmachung an einen abgegrenzten Personenkreis für spezifische Zwecke zulässig.[105] Die Aufbewahrung von Daten bei der Forschungseinrichtung selbst ist nunmehr nach § 60d Abs. 5 UrhG möglich, solange sie für den Zweck der wissenschaftlichen Forschung oder für die Überprüfung wissenschaftlicher Erkenntnisse erforderlich ist.

Vertiefende Literaturhinweise: *De la Durantaye/Raue*, Urheberrecht und Zugang in einer digitalen Welt, Recht und Zugang (RuZ) 2020, 83; *Specht*, Die neue Schrankenregelung für Text

---

[100] Hierzu näher siehe *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6.

[101] *Raue*, ZUM 2021, 793, 799; die genaue Ausgestaltung angemessener Sicherheitsvorkehrungen ist zum jetzigen Zeitpunkt nicht geklärt.

[102] *Raue*, ZUM 2021, 793, 799; zu solchen vertrauenswürdigen Stellen siehe Erwägungsgrund 15 S. 3 der 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 (DSM-RL).

[103] *Raue*, ZUM 2021, 793, 799 mit Blick auf: DFG, Leitlinien zur Sicherung guter wissenschaftlicher Praxis, Kodex, 2019, Leitlinie 12, 13, insbesondere 17; Argumente für eine längere Aufbewahrungszeit: *Heesen/Jüngels*, RuZ 2021, 45, 50.

[104] *Raue*, ZUM 2021, 793, 799: Forschende haben hier eine Einschätzungsprärogative, die gerichtlich lediglich eingeschränkt auf Missbrauch überprüft werden kann; dazu auch *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6.

[105] Zum Begriff der „öffentlichen Zugänglichmachung" in § 60d Abs. 4 UrhG siehe *Erler-Fridgen*, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6, IV.

und Data Mining und ihre Bedeutung für die Wissenschaft, Ordnung der Wissenschaft (OdW) 2018, 285; *Raue*, Die Freistellung von Datenanalysen durch die neuen Text und Data Mining-Schranken (§§ 44b, 60d UrhG), Zeitschrift für Urheber- und Medienrecht (ZUM) 2021, 793.

VII. Literaturverzeichnis

Hartwig Ahlberg, Horst-Peter Götting, Anne Lauber-Rönsberg (Hrsg.), Beck'scher Onlinekommentar Urheberrecht, 33. Edition 2022, C.H. Beck München.

David M. Blei, Probabilistic Topic Models, http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf (abgerufen am 12.01.2022).

Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde, Hagen Langer (Hrsg.), Computerlinguistik und Sprachtechnologie, 3. Aufl. 2010, Spektrum Akademischer Verlag, Heidelberg.

Thomas Dreier, Gernot Schulze (Hrsg.), Urheberrechtsgesetz, 7. Aufl. 2022, C. H. Beck München.

Katharina Erler-Fridgen, Die Nutzung wissenschaftlicher Ausgaben für Textanalysen, IRDT PAPER*SERIES* Nr. 1.

Katharina Erler-Fridgen, Kriterien der urheberrechtlichen Schutzfähigkeit von Texten und Sammelwerken, IRDT PAPER*SERIES* Nr. 2.

Katharina Erler-Fridgen, Die Präsentation von Textteilen als Ergänzung von Textanalysen, IRDT PAPER*SERIES* Nr. 3.

Katharina Erler-Fridgen, Die Text und Data Mining-Schranken und ihr Rahmen für Textanalysen in den Digital Humanities, IRDT PAPER*SERIES* Nr. 6.

Katharina Erler-Fridgen, Datenbanken als Quelle oder Ergebnis von Textanalysen – Datenbankwerkschutz und das Leistungsschutzrecht des Datenbankherstellers, IRDT PAPER*SERIES* Nr. 4.

Erler-Fridgen, Das Zitat und dessen Rahmen für Belege bei Textanalysen, IRDT PAPER*SERIES* Nr. 7.

Katharina de la Durantaye, Benjamin Raue, Urheberrecht und Zugang in einer digitalen Welt, Urheberrechtliche Fragestellungen des Zugangs für Gedächtnisinstitutionen und die Digital Humanities, Recht und Zugang (RuZ) 2020, 83.

Stefan Engelberg, Kookkurrenzanalyse, Linguistische Methodenlehre, FS 2009, Uni Mannheim, https://www1.ids-mannheim.de/fileadmin/lexik/lehre/engelberg/Webseite_LingMeth/Skript_05.pdf. (abgerufen am 12.01.2022).

Thomas Hoeren, Ulrich Sieber, Bernd Holznagel, Handbuch Multimedia-Recht, 57. EL 2021, C.H.Beck München.

Martin Huber, Sybille Krämer (Hrsg), Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 3), http://zfdg.de/sb003_002 (abgerufen am 12.01.2022).

Gabe Ignatow, Rada Mihalcea, Text Mining – A Guidebook for the Social Sciences, 2017, Sage Publications Thousand Oaks/London/New Dehli/Singapore.

Fotis Jannidis, Hubertus Kohle, Malte Rehbein (Hrsg.), Digital Humanities – Eine Einführung, 2017, J.B. Metzler Verlag, Stuttgart.

Florian Jotzo, Der Schutz großer Textbestände nach dem UrhG, Recht und Zugang (RuZ) 2020, 128.

Felicitas Kleinkopf, Janina Jacke, Markus Gärtner, Text- und Data-Mining, Urheberrechtliche Grenzen der Nachnutzung wissenschaftlicher Korpora bei computergestützten Verfahren und digitalen Ressourcen, Zeitschrift für IT-Recht und Recht der Digitalisierung (MMR) 2021, 196.

Claudia Kunze, Lothar Lemnitzer, Computerlexikographie. Eine Einführung, 2007, Gunter Narr Verlag, Tübingen.

Georg Rehm, Andreas Witt, Lothar Lemnitzer (Hrsg.), Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Proceedings of the Biennial GLDV Conference 2007, Narr 2007, Tübingen.

Ulrich Loewenheim, Matthias Leistner, Ansgar Ohly (Hrsg.), Schricker/Loewenheim, Urheberrecht, 6. Auflage 2020, C.H. Beck München.

Benjamin Raue, Das Urheberrecht der digitalen Wissen(schaft)sgesellschaft, Gewerblicher Rechtsschutz und Urheberrecht (GRUR) 2017, 11.

Benjamin Raue, Rechtssicherheit für datengestützte Forschung, Die Text-und-Data-Mining-Schranken in Art. 3 und 4 DSM-RL, Zeitschrift für Urheber- und Medienrecht (ZUM) 2019, 684.

Benjamin Raue, Die Freistellung von Datenanalysen durch die neuen Text und Data Mining-Schranken (§§ 44b, 60d UrhG), Zeitschrift für Urheber- und Medienrecht (ZUM) 2021, 793.

Benjamin Raue, Free Flow of Data? The Friction Between the Commission's European Data Economy Initiative and the Proposed Directive on Copyright in the Digital Single Market, International Review of Intellectual Property and Competition Law (IIC) 2018, 379.

Benjamin Raue, Der schleichende Tod des Bearbeitungsrechts – Vervielfältigung, Bearbeitung, Pastiche und freie Benutzung im Urheberrecht, Archiv für Presserecht (AfP) 2022, 1.

Leif Scheuermann, Die Abgrenzung der digitalen Geisteswissenschaften, in Digital Classics Online 2 (2016), 1, S. 61, https://journals.ub.uni-heidelberg.de/index.php/dco/article/view/22746/21865 (abgerufen am 12.01.2022).

Natascha Schumann, Einführung in die digitale Langzeitarchivierung, S. 44 https://www.ssoar.info/ssoar/bitstream/handle/document/45740/ssoar-2012-schumann-Einfuhrung_in_die_digitale_Langzeitarchivierung.pdf (abgerufen am 12.01.2022).

Ray Siemens, John Unsworth, Susan Schreibman, A Companion to Digital Humanities, Blackwell Publishing, http://www.digitalhumanities.org/companion/ (abgerufen am 12.01.2022).

Louisa Specht, Die neue Schrankenregelung für Text und Data Mining und ihre Bedeutung für die Wissenschaft, Ordnung der Wissenschaft (OdW) 2018, 285.

Gerald Spindler, Text und Data Mining – urheber- und datenschutzrechtliche Fragen, Gewerblicher Rechtsschutz und Urheberrecht (GRUR) 2016, 1112.

Artur-Axel Wandtke, Winfried Bullinger (Hrsg.), Praxiskommentar Urheberrecht, 5. Aufl. 2019, C.H. Beck München.