

Einleitung

§ 1 Problemaufriss

Mit computerlinguistischen Mitteln wurde *Joanne K. Rowling* im Jahr 2013 als wahre Autorin der Romane enttarnt, die sie unter ihrem Pseudonym *Robert Galbraith* veröffentlichte.¹ Mithilfe der Software „Deep-L“² ist es möglich, Texte aus 26 Sprachen in einer erstaunlichen Qualität übersetzen, die von der Vereinigung *OpenAI* entwickelte Software GPT-3 vermag es, ganze Texte generieren³ und das „Datenleak“ der sog. „Panama Papers“ konnte zu investigativjournalistischen Zwecken nach Inhalten und Zusammenhängen durchsuchbar gemacht werden.⁴ Das sind nur einige wenige Beispiele, die gemeinsam haben, dass sie darauf basieren, dass Computer lernen, die natürliche Sprache zu verstehen und unmittelbare inhaltliche Schlussfolgerungen als Text- oder Datengrundlagen ziehen. Letztlich basieren alle diese Beispiele auf der Technologie des sog. Text- und Data-Minings.

Das Text- und Data-Mining, im Folgenden als TDM abgekürzt, ist ein digitales Analyseverfahren, bei dem Datensätze, bestehend aus (teil-)strukturierten Rohdaten oder Texten, aber auch Bildern oder Musikstücke auf unterschiedliche Muster bzw. Zusammenhänge untersucht werden. Begrifflich bezieht sich der englische Begriff „to mine“ ursprünglich auf den Prozess des Abbauens von Bodenschätzen.⁵ Im übertragenen Sinne geschieht dies auch beim TDM: Aus digitalen Ressourcen werden Erkenntnisse gewonnen.⁶ Dafür ist in der Regel erforderlich, dass die Daten zu einem einheitlichen Datensatz vorverarbeitet wurden, der das sog. Korpus bildet. Nur dann sind die den Datensätzen immanenten Muster für diese Algorithmen sichtbar. Dazu werden umfangreiche Arbeitsschritte vorge-

1 *B. Zimmer*, Rowling and "Galbraith": an authorial analysis (16.07.2013).

2 <https://www.deepl.com/translator>.

3 *Lobe*, Frankfurter Allgemeine Sonntagszeitung 20.09.2020, S. 36.

4 *Munzinger*, Panama und Paradise Papers (21.06.2018); *Mukhopadhyay/Ghosh*, SSRN Journal 2020.

5 So die Oxford English Learner's Dictionaries sowie das PONS Wörterbuch, <https://www.oxfordlearnersdictionaries.com/definition/english/mining>; <https://de.pons.com/%C3%BCbersetzung/englisch-deutsch/mine>.

6 *Schölkopf*, Causality for Machine Learning (24.11.2019), S. 1 f.

nommen, in denen die Texte oder Daten, die analysiert werden sollen, vereinheitlicht und mit weiteren Informationen, die auf das konkrete Analyseziel angepasst sind, angereichert werden. Die Technologie des TDMs kann also als Forschungsmethode, d. h. als Werkzeug zur Beantwortung wissenschaftlicher Fragestellungen, eingesetzt werden. Weiter ist die Technologie in vielen Fällen Grundlage von Systemen, die zur sog. Künstlichen Intelligenz, abgekürzt als KI, zählen.⁷ Insgesamt hat die Technologie des TDMs eine große Bedeutung für Innovation und Wissenschaft. Nicht ohne Grund bezeichnet die DSM-RL aus dem Jahr 2019 das Verfahren als die „*vorherrschende Technik der Digitalwirtschaft*“.⁸

Wie viele andere digitale Prozesse auch ist das TDM wesentlich vom Urheberrecht reglementiert, obwohl die rechtlichen Probleme, die das TDM aufwirft, lange vorwiegend aus datenschutzrechtlicher Hinsicht diskutiert wurden⁹ und urheberrechtliche Aspekte erst nach 2010 Aufmerksamkeit erfuhren.¹⁰ Wenn also Ressourcen Gegenstand von TDM-Analysen werden sollen, die durch das UrhG geschützt sind, stellt sich die Frage, ob und in welchem Umfang ihre Nutzung zulässig ist. Das gilt sowohl für die algorithmusbasierte Analyse als auch für die Aufbereitung der genutzten Ressourcen. Mittels spezifischer Schrankenbestimmungen, von denen die erste auf deutscher Ebene schon vor dem Tätigwerden auf EU-Ebene geschaffen wurde (§ 60d UrhG a. F.) und die in Umsetzung der Vorgaben der späteren Art. 3 und 4 der DSM-RL reformiert wurde (§§ 44b, 60d UrhG), sollte das TDM auch an urheberrechtlichen Schutzgegenständen ermöglicht werden.¹¹ Insbesondere die DSM-RL zeigt dabei eine wertungsmä-

7 *Truyens/van Eecke*, CLSR 2014, S. 153, 154, die darüber hinaus auf die engen Verknüpfungen zum Datenbank-Management verweisen.

8 DSM-RL Erw. 8.

9 *Truyens/van Eecke* in: European Language Resources Association (Hrsg.), Proceedings of the Ninth International Conference on Language Resources and Evaluation ({LREC}'14), S. 2182, 2182.

10 *Truyens/van Eecke* in: European Language Resources Association (Hrsg.), Proceedings of the Ninth International Conference on Language Resources and Evaluation ({LREC}'14), S. 2182, 2182; Europäische Kommission, Mitteilung der Europäischen Kommission über Inhalte im digitalen Binnenmarkt, COM(2012) 789 final (18.12.2012), S. 4.

11 Die Gesetzesbegründung des UrhWissG enthält hierzu die Angabe: „Die Vorschrift ermöglicht es, auf gesetzlicher Grundlage Werke mit Inhalten aller Art automatisiert auszuwerten, z. B. Werke mit Texten, Daten, Bildern, Tönen oder audiovisuellen Inhalten, um damit nicht kommerzielle wissenschaftliche Forschung zu betreiben. Schlagwortartig wird dieser Vorgang häufig als sogenanntes Text und Data Mining bezeichnet. Die Regelung hat keinen Vorläufer im bisherigen deutschen Recht.“, BT-Drs. 18/12329, S. 40; die DSM-RL will über die zuvor bestehenden unionsrechtlichen

ge Privilegierung des TDMs, indem sie die Erlaubnisse des TDMs für die Mitgliedsstaaten verpflichtend ausgestaltet und auch das innovationspolitische Potenzial des TDMs würdigt.¹²

Ob es auf Grundlage der verschiedenen Normfassungen – insbesondere auf Grundlage der aktuellen Rechtslage – möglich ist, das Potenzial des TDMs auch bei der Analyse urheberrechtlicher Schutzgegenstände zu nutzen, bedarf einer genauen Analyse, denn die Regulierung digitaler Prozesse wirft in einem Urheberrecht, das weitestgehend einem analogen Zeitalter entstammt, verschiedene Probleme auf. In Bezug auf das TDM äußern sich diese Probleme etwa bei der Beantwortung der Frage danach, ob das TDM mit dem Werkgenuss vergleichbar ist, der traditionell urheberrechtlich freigestellt ist, für den aber in digitalen Umgebungen andere Mechanismen gelten. Gerade im Bereich der wissenschaftlichen Forschung sind außerdem besondere Anforderungen an einen Rechtsrahmen zu beachten, die die Einhaltbarkeit der wissenschaftlichen Eigengesetzlichkeiten, z. B. in Gestalt der Anforderungen der guten wissenschaftlichen Praxis,¹³ aber auch das Erfordernis von Rechtssicherheit, betreffen. Diese wissenschaftlichen Eigengesetzlichkeiten beziehen sich nicht nur auf die Beachtung wissenschaftlicher Grundsätze während des Forschungsprozess als solchen, sondern erfordern insbesondere eine Transparenz der wissenschaftlichen Forschung, d. h. auch der Forschungsergebnisse.

Letztlich verfolgt die Arbeit das Ziel, die Rechtsentwicklung auf ihre Eignung zur Regulierung des technischen Prozesses des TDMs zu überprüfen, Defizite aufzuzeigen und basierend auf einer teleologischen, systematischen sowie interdisziplinären Argumentation gezielte Rechtsanpassungen vorzuschlagen.

Möglichkeiten insbesondere rechtssichere Bedingungen schaffen: *„Die Rechtsunsicherheit im Hinblick auf Text und Data Mining sollte beseitigt werden, indem für Hochschulen und andere Forschungsorganisationen sowie für Einrichtungen des Kulturerbes eine verbindliche Ausnahme für das ausschließliche Recht auf Vervielfältigung, aber auch auf das Recht, Entnahmen aus einer Datenbank zu untersagen, eingeführt wird.“*, DSM-RL Erw. 11 S. 1.

- 12 So betont sie in Erw. 18 S. 1: *„Verfahren des Text und Data Mining haben nicht nur im Zusammenhang mit der wissenschaftlichen Forschung hohe Bedeutung, sondern sie werden auch in großem Umfang sowohl von privaten als auch öffentlichen Einrichtungen eingesetzt, um große Datenmengen in verschiedenen Lebensbereichen und zu unterschiedlichen Zwecken zu analysieren, auch für staatliche Dienste, komplexe unternehmerische Entscheidungen und die Entwicklung neuer Anwendungen oder Technologien.“*
- 13 Z. B. kodifiziert in Deutsche Forschungsgemeinschaft, Leitlinien zur Sicherung guter wissenschaftlicher Praxis (2019).

§ 2 Herangehensweise

A. Grenzen der Untersuchung

Bei der urheberrechtlichen Betrachtung des TDMs kommt eine Vielzahl an Einzelfragen auf, die eine Eingrenzung des Untersuchungsgegenstands erfordert. Die vorliegende Untersuchung beschränkt sich auf die Analyse des Rechtsrahmens in Bezug auf das TDM bei der Nutzung von Schriftwerken, § 2 Abs. 1 Nr. 1, Abs. 2 UrhG, Datenbankwerken, §§ 4 Abs. 2, 2 Abs. 2 UrhG, sowie von Datenbanken, § 87a UrhG, die als Leistungsschutzrechte ins UrhG aufgenommen wurden, für das TDM. Nicht untersucht werden hingegen Besonderheiten, die sich bei der Analyse von Werken der bildenden Kunst oder Musik ergeben.¹⁴ In Bezug auf das TDM in Web-Kontexten, das durch Web-Crawler oder -Scraper durchgeführt bzw. mittels jener praktiziert wird, ergaben sich seit den 2000er-Jahren zahlreiche Gerichtsverfahren in kommerziellen Kontexten, in denen zumeist Gegenstand war, dass Webrobots Daten aus Webseiten abriefen und in Meta-Datenbanken einpflegten. Diese Fragestellungen werden nur referenziell zur urheberrechtlichen Betrachtung des TDMs thematisiert, nicht aber hinsichtlich eigener Rechtsfragen, die durch sie aufgeworfen werden.¹⁵ Nicht weiter analysiert werden Fragen nach einem Schutz der Ausgabe aus dem Bereich des maschinellen Lernens, z. B. sog. generativer Modelle, die vermeintliche Werke oder Schutzgegenstände, z. B. in einem bestimmten künstlerischen Stil erschaffen können.¹⁶ Ebenso wenig wird thematisiert, ob und inwieweit die Konzeption von Forschungsfragen urheberrechtlich schutzfähig ist.

14 Vgl. dazu allerdings *Rack/Frieler* in: Gräfe/Telemedicus e.V. (Hrsg.), Tagungsband zur Sommerkonferenz 2019: Über den Tellerrand, S. 27.

15 Thematisiert wurde in diesem Kontext in erster Linie die Frage nach dem Bestehen eines virtuellen Hausrechts und die Bezüge zur Informationsfreiheit, vgl. *v. Schönfeld*, Screen Scraping und Informationsfreiheit, *A. Schmidt*, Virtuelles Hausrecht und Webrobots; *Elteste*, CR 2015, S. 447; *Gausling*, CR 2021, S. 609; *Golla/v. Schönfeld*, K&R 2019, S. 15 ff.; *Schapiro/Ždanowiecki*, MMR 2015, S. 497 ff.

16 Das untersuchen *Käde*, Kreative Maschinen und Urheberrecht, S. 173 ff.; *Grätz*, Künstliche Intelligenz im Urheberrecht, S. 45 ff.; *Dornis*, GRUR 2019, S. 1252, 1256 ff.; *Ebinger/Stiemerling*, CR 2018, S. 761 ff.; *Hetmank/Lauber-Rönsberg*, GRUR 2018, S. 574 ff.; *Lauber-Rönsberg*, GRUR 2019, S. 244 ff. Im Bereich sprachbasierter Werke ist es z. B. möglich, automatische Abstracts zu erstellen oder Texte im Stil bestimmter Autoren verfassen zu lassen, zu Textgeneratoren *Gräfe/J. Kahl*, MMR 2021, S. 121 ff.; *Lobe*, Frankfurter Allgemeine Sonntagszeitung 20.09.2020, S. 36.

B. Interdisziplinärer Ansatz

Die Arbeit untersucht das TDM als Methode der wissenschaftlichen Forschung und der Datenanalyse anhand urheberrechtlicher Schutzgegenstände. Zentrale Untersuchungsgegenstände bilden deswegen das urheberrechtliche Regelungsregime im europäischen Mehrebenensystem, das Verhältnis von Schutzrecht und Schranke sowie die Besonderheiten des Wissenschaftsurheberrechts. Daneben werden auch grundrechtliche Wertungen miteinbezogen, die einerseits das Verhältnis des Schutzes des Eigentums zu Inhalts- und Schrankenbestimmungen, Art. 14 Abs. 1 S. 1, 2 GG und andererseits den Einfluss der Freiheit von Wissenschaft und Forschung auf urheberrechtliche Schrankenbestimmungen, Art. 5 Abs. 3 S. 1 Alt. 2 GG, betreffen.

Daneben verfolgt die Arbeit einen interdisziplinären Ansatz, bei dem Grundprinzipien und Erkenntnisse anderer Fachdisziplinen für die juristische Betrachtung nutzbar gemacht werden: Zur Erfassung des technischen Verfahrens wird in größerem Umfang informationstechnische Literatur gewürdigt. Weiter wird die Bedeutung des TDM-Verfahrens und der Forschungsdaten des TDMs – der Korpora – aus wissenschaftstheoretischer und kulturwissenschaftlicher Perspektive betrachtet. Zur Erfassung der ökonomischen Auswirkung des Urheberrechts auf Innovation bedarf es außerdem der Berücksichtigung rechtsökonomischer Erwägungen eingeschlossen der Auswertung einiger weniger empirischer Studien, die diese Problematik konkret in Bezug auf das TDM untersuchen.

C. Gang der Untersuchung

Der erste Teil der Arbeit bildet die Grundlage der Untersuchung und stellt dar, was das TDM ist und welche Bedeutung es als Forschungsmethode, für das maschinelle Lernen sowie weitere Teilbereiche der künstlichen Intelligenz hat. Anschließend nähert sich die Untersuchung einem Ablauf des TDMs an, der in gewissem Umfang auf verschiedene Anwendungen des TDMs übertragen werden kann. Herausgearbeitet wird daneben, wie die Muster einzuordnen sind, die das TDM hervorbringt und worin ihr besonderes Potenzial liegt. Im weiteren Fortgang widmet sich der erste Teil der Arbeit vorwiegend der Steuerungswirkung von wissenschaftlicher Forschung sowie Innovation durch das Urheberrecht, die später dazu nutzbar gemacht wird, den Rechtsrahmen für das TDM zu beurteilen.

Der zweite Teil der Arbeit geht dem Forschungs- und Analyseprozess des TDMs in seiner urheberrechtlichen Einordnung und der Frage, ob es aus übergeordneten Gründen freigestellt oder jedenfalls freizustellen ist, nach. Auf der Analyse des Urheberrechtsrahmens des TDM-Prozesses aufbauend wird die Rechtsentwicklung in Bezug auf ihre Eignung und Vollständigkeit analysiert.

Im dritten Teil stehen die Abläufe im Vordergrund, die im Anschluss an die eigentlichen Forschungsarbeiten stattfinden – die Möglichkeit wissenschaftlichen Überprüfung durch eine transparente Archivierung und Zugänglichkeit für die Überprüfungen und zur weiteren Nutzung der Korpora für Anschlussforschungen, über deren Definition, Unterteilung und Bezeichnung weitestgehend Uneinigkeit herrscht. Aus den Rückschlüssen, die diesbezüglich aus der Wissenschaftsfreiheit sowie aus der Wissenschaftstheorie und der Kulturwissenschaft gezogen werden, wird ein Maßstab eines geeigneten Rechtsrahmens gebildet. Schließlich wird das geltende Recht daraufhin überprüft, inwieweit es die gesetzten Maßstäbe erfüllt. Alternativ werden Ansätze vorgestellt, wie Korpora über den explizit normierten Umfang hinaus nachnutzbar gemacht werden können. Dabei stehen sich schließlich technische und rechtliche Ansätze gegenüber, die gleichermaßen relevant in technischen Sachverhalten und ihrer urheberrechtlichen Regulierung sind.

Im vierten Teil der Arbeit werden die Erkenntnisse aus der Arbeit in Thesenform zusammengefasst, die schließlich in konkreten Verbesserungsvorschlägen münden. Zuletzt wagt die Arbeit einen Ausblick.