

# TAGUNGSBERICHT

Katharina Erler\*

## DFG-Expertenworkshop: Strategien für die Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte

### I. Zur Tagung

Am 27.11. und 28.11.2019 sowie am 16.01. und 17.1.2020 fand an der Universität Trier die von der Deutschen Forschungsgemeinschaft geförderte Workshop-Reihe „Strategien für die Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte“ statt. Auf Einladung des Instituts für Recht und Digitalisierung Trier (IRDT) und des Trier Center for Digital Humanities (TCDH) diskutierten Experten der Rechtswissenschaften, der Digital Humanities und der Informatik gemeinsam mit Vertretern aus dem Bibliothekssektor, wie urheberrechtlich geschützte Textbestände für die Forschung durch Dritte nutzbar gemacht werden können.

### II. Zu den Inhalten

Kern der Diskussionen war die Frage, ob und in welcher Form urheberrechtlich geschützte Texte in abgeleitete Textformate transformiert werden können, um diese in eine nutzbare Quelle für Analyseverfahren der Digital Humanities zu überführen. Dahinter stand die These, dass abgeleitete Textformate Dritten ermöglichen, Forschungsdaten nachzunutzen und Forschungsergebnisse zu reproduzieren, ohne dafür auf die Freistellung des Text-und-Data-Mining für die wissenschaftliche Forschung in Art. 3 der Richtlinie 2019/790 (DSM-RL)<sup>1</sup> und § 60d UrhG<sup>2</sup> zurückgreifen zu müssen. Neben der urheberrechtlichen Bewertung solcher Transformationen wurden die abgelei-

\* Die Verfasserin Dipl.-Jur. Katharina Erler ist wissenschaftliche Mitarbeiterin am Institut für Recht und Digitalisierung Trier (IRDT, Universität Trier) und Doktorandin der Forschungsstelle für Verbraucherrecht (FfV, Universität Bayreuth).

1 Richtlinie 2019/790 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt (DSM-Richtlinie).

2 Eingeführt mit dem Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (UrhWissG).

teten Textformate zugleich an den Anforderungen verschiedener Analyseszenarien der Digital Humanities gemessen. Auch Entwicklungspotentiale der Transformation und Analyse der Texte wurden in den Blick genommen, die der Einsatz von Künstlicher Intelligenz ermöglicht. Begleitet wurden die Diskussionen von Experten aus dem Bibliothekssektor, die ihre Perspektive einbrachten, welche Rolle die Bibliotheken bei dem Prozess spielen können, insbesondere, wie Analyseverfahren umgesetzt werden können. Die Diskussionsergebnisse konnten schließlich direkt mit den Erfahrungen aus einer bereits erfolgten Umsetzung in den USA (Projekt HathiTrust Research Center) zusammengeführt werden.

Zur Eröffnung des ersten Teils der Workshops-Reihe sprach *Prof. Dr. Michael Jäckel*, Präsident der Universität Trier, über die Herausforderungen der Digitalisierung und dass eine neue Sensibilität für die Datennutzung in der Forschung erforderlich sei. Mit Blick auf das Humboldt'sche Ziel des globalen Wissensaustauschs hob er die Bedeutung des Workshops hervor. Denn neue Erkenntnisse in der Sekundärforschung könnten einfacher dadurch gewonnen werden, dass Primärdaten systematisch zugänglich gemacht werden.

*Prof. Dr. Christof Schöch* (Ko-Direktor des Trier Center for Digital Humanities) und *Prof. Dr. Benjamin Raue* (Geschäftsführender Direktor des Instituts für Recht und Digitalisierung Trier) führten hiernach in die Zielsetzung des DFG-Expertenworkshops ein. Der Workshop solle eine Plattform bieten, um im interdisziplinären Austausch gemeinsam Strategien für die Umsetzung der Analyse urheberrechtlich geschützter Texte zu entwickeln. Dabei sei es entscheidend, die vielgestaltigen Anforderungen aller eingebrachten Perspektiven – des Urheberrechts, der Digital Humanities und Informatik sowie die der Experten aus dem Bibliothekssektor – zu berücksichtigen.

Ziel des ersten Teils der Workshop-Reihe war es, das Konzept für eine Handreichung („White Paper“) zu entwickeln, welche im Nachgang die Diskussionsergebnisse dokumentieren und interdisziplinär aufbereiten soll. Als Ausgangspunkt der Entwicklung eines gemeinsamen Konzepts führten in einem ersten Schritt *Dr. Florian Jotzo* (Universität Kiel) und *Dr. Karina Grisse* (Universität zu Köln) in die aktuellen rechtlichen Rahmenbedingungen ein. Um sich der Evaluation urheberrechtlicher Implikationen der Transformation von Texten anzunähern, ging *Jotzo* insbesondere auf die Voraussetzungen des urheberrechtlichen Schutzes von Texten und Textteilen ein. Er betonte zum einen, dass die Rechtsprechung an den schutzauslösenden Faktor der Gestaltungshöhe je nach vorliegender Textart unterschiedliche Maßstäbe anlegen würde. Zum anderen identifizierte er drei wesentliche Faktoren für die Schutzfähigkeit: die Länge eines Textes oder Textteils, das Vorliegen von kreativ gestaltetem Inhalt und die Form des Textes. Diese zöge die Rechtsprechung in einer Gesamtbetrachtung zur Beurteilung der Schutzfähigkeit heran. Darauf aufbauend widmete sich *Grisse* der Frage, wie der Urheberrechtsschutz durch die Transformation von Texten beseitigt werden könnte. Dadurch könne gegebenenfalls ein höheres Maß an Rechtssicherheit für automatisierte Analysen von Texten in der Forschung erzielt werden. Als eine mögliche Anregung für die urheberrechtliche Evaluation abgeleiteter Textformate zog sie unter

anderem eine im Jahr 2019 ergangene Entscheidung des EuGH<sup>3</sup> zum Sampling fremder Audiofragmente heran. In dieser Entscheidung hatte der EuGH bei fehlender Wiedererkennbarkeit des Werkes keine Vervielfältigung der Audiofragmente angenommen.

Im zweiten Schritt stellten *Prof. Dr. Achim Rettinger* (Universität Trier) und *Prof. Dr. Christof Schöch* (Universität Trier) die Anforderungen der Anwendungswissenschaften an abgeleitete Textformate dar, welche durch die Transformation der Volltexte generiert werden könnten. *Rettinger* erläuterte Verfahren des maschinellen Lernens, die in der Computerlinguistik angewendet werden. Dabei erläuterte er mögliche Entwicklungspotentiale für die Analyse von Texten durch das Verfahren des Word Embedding. Anschließend stellte *Schöch* die zwei aus Anwendersicht wesentlichen Leitfragen der Transformation von Textbeständen und hieraus abgeleiteter Textformaten dar. Für geisteswissenschaftliche Anwendungskontexte sei zum einen entscheidend, welche Analysemethoden mit abgeleiteten Formaten offenstehen. Zum anderen sei für die Bewertung der unterschiedlichen Formate wesentlich, welche Informationen für weitere Analysen notwendig und welche weiteren Metadaten für die Nutzenden enthalten sind. Um diese Anforderungen an abgeleitete Formate deutlich zu machen, gab er einen Überblick über die verschiedenen Methoden und Analyseverfahren des Text- und Data-Mining und maschinellen Lernens. Auch erläuterte er deren Einsatz zur geisteswissenschaftlichen Analyse von Textsammlungen.

Im dritten Schritt schilderten Vertreter aus dem Bibliothekssektor, *PD Dr. phil. Frédéric Döhl*, *Ass. iur.* (Deutsche Nationalbibliothek Leipzig), *Dr. Peter Leinen* (Deutsche Nationalbibliothek Frankfurt a. M.) und *Jörg Röpke* (Universitätsbibliothek Trier), welche Herausforderungen bei der Umsetzung von Text- und Data-Mining-Verfahren aus Perspektive der Infrastruktureinrichtungen bestünden. Art. 3 DSM-RL privilegiert nun Einrichtungen des Kulturerbes. Bibliotheken müssten sich dessen Anforderungen an die rechtliche und technische Umsetzung stellen. Insbesondere wurde der Servicecharakter von Text- und Data-Mining („Text- und Data-Mining as a Service“) bei Umsetzung der Verfahren durch die Bibliotheken diskutiert.

Im vierten Schritt präsentierte *Prof. Peter Organisciak*, *Ph.D.* (University of Denver) das Projekt HathiTrust Research Center, das gemeinsam von der Indiana University, University of Illinois und HathiTrust getragen wird. Dieses Research Center stellt Forschern Software und Services zur Verfügung, um Text- und Data-Mining-Verfahren auf Basis der HathiTrust Textsammlung durchzuführen. Er sprach über die rechtlichen Herausforderungen des Projekts mit Blick auf das US-amerikanische IP law,<sup>4</sup> beschrieb die Aufbereitung der Textbestände für die Analyse und präsentierte die derzeit

<sup>3</sup> EuGH, Urt. v. 29.7.2019 – C-476/17, ECLI:EU:C:2019:624 = MMR 2019, 596 – *Pelham u.a.*

<sup>4</sup> In *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) klassifizierte das Gericht die von HathiTrust ermöglichte „use of full-text search“ und „methods such as text mining“ als *fair use* unter 17 U.S.C. § 107.

als Services ausgestalteten verschiedenen Nutzungs- und Zugriffsmöglichkeiten für Forschende.

Als Zwischenergebnis des ersten Teils der Workshops-Reihe wurde das Konzept einer Handreichung zu abgeleiteten Textformaten sowie der Schwerpunkt des zweiten Workshops präzisiert. Dabei wurde der Fokus des zweiten Workshops auf die Evaluation einzelner Beispiele für abgeleitete Textformate gesetzt, wobei alle beteiligten Perspektiven einfließen sollten.

Anknüpfend an die Ergebnisse und Diskussionen des ersten Workshops, wurde zu Beginn des zweiten Workshops ein Überblick über die urheberrechtlichen Maßstäbe zur Bewertung abgeleiteter Textformate gegeben. Das nunmehr präzisierte Prinzip abgeleiteter Textformate wurde anschließend aus Sicht der Digital Humanities vorgestellt. Schließlich wurden einzelne Textformate aus den verschiedenen Perspektiven der Rechtswissenschaft, der Digital Humanities und Informatik sowie aus der Perspektive des Bibliothekssektors evaluiert.

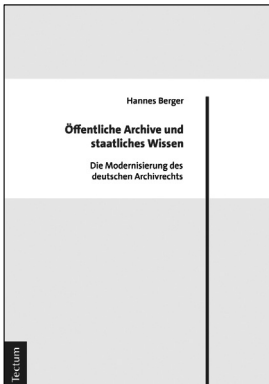
*Jotzo* führte dazu im ersten Schritt die Essenz der relevanten Kriterien für die Begründung urheberrechtlichen Schutzes – Textart, kreativ gestalteter Inhalt, Form und Länge des Textes – zusammen. Auf dieser Basis könnten zwei Orientierungslinien für die Evaluation abgeleiteter Textformate genutzt werden: einerseits die Extraktion schutzfreier Elemente (bloßer Informationen) und andererseits die Beseitigung schutzbegründender Merkmale (kreativer Merkmale). *Grisse* betonte, dass das Kriterium der Erkennbarkeit des Werkes entscheidend sei für den Weg aus dem Urheberrechtsschutz, der durch Transformation der Texte erreicht werden könne. Sie erörterte den Maßstab der Erkennbarkeit insbesondere mit Blick auf Werke, deren stilistische Eigenheiten besonders stark ausgeprägt sind. Ob ein Werk durch Transformation unkenntlich gemacht worden oder immer noch rekonstruierbar sei, müsse durch ein Bündel an möglichen Kriterien anhand einer Einzelfallbetrachtung bestimmt werden. Zugleich könne sich die Vervielfältigungshandlung bei Rekonstruktion des Werkes auf den Rekonstruierenden und damit auf eine nachgelagerte Ebene verschieben.

Nach diesen rechtlichen Ausführungen führte *Schöch* seine weiteren Überlegungen zu abgeleiteten Textformaten aus Anwendersicht aus. Sie sollten drei Kriterien genügen: Neben einer möglichst hohen Informationsdichte und breiten Einsatzmöglichkeit für Analyseverfahren sollten diese auch aus urheberrechtlicher Sicht unbedenklich sein. Zugleich solle die Fokussierung auf wenige Formatvorschläge Kooperationen verschiedener Institutionen ermöglichen. Er stellte die Parameter der Formatgestaltung – etwa die Informationsselektion, die Sequenzinformation, die Positionsangabe und Häufigkeitsinformationen – vor und präsentierte eine Auswahl an Formaten und deren konkrete Anwendung auf Textbeispiele.

Diese boten im Anschluss eine Grundlage für eine interaktive Diskussion, in der die Textformate im Einzelnen evaluiert wurden. *Rettinger* wies dabei auf das mögliche Potential des Verfahrens des Word Embedding hin. Vorteil dieses Verfahrens wäre, dass es die Rekonstruierbarkeit des Werkes jedenfalls erheblich erschwere. Da aus diesem Grund die Sequenzinformationen des Werkes erhalten bleiben könnten, mache es eine Sentiment-Analyse der Texte möglich. Aus diesem Anlass wurde für alle abgeleiteten

Textformate diskutiert, mit welchen empirischen und mathematisch-analytischen Ansätzen die Rekonstruierbarkeit der Volltexte und des damit verbundenen Rechenaufwands erforscht werden könnten. Auch das Kriterium der Erkennbarkeit eines transformierten Werkes durch den durchschnittlichen Rezipienten könne auf eine empirische Basis gestellt werden. Schließlich wurden Fragen der Standardisierung von abgeleiteten Textformaten erörtert. Ansatzpunkt dieser Diskussion war, dass verlässliche Standards Kooperationen mit anderen Forschungseinrichtungen und die Zusammenarbeit zwischen einzelnen Infrastruktureinrichtungen erleichtern könnten.

Die Ergebnisse der Workshop-Reihe sollen in einer Handreichung („White Paper“) veröffentlicht werden. In der Handreichung werden die Perspektiven der Rechtswissenschaft, der Digital Humanities und Informatik sowie die Erfahrungen aus dem Bibliothekssektor zusammengeführt. Auch weiterführende Projekte zur Evaluierung abgeleiteter Textformate sind vorgesehen. Beispielsweise planen die Beteiligten, die Reproduzierbarkeit abgeleiteter Textformate sowie die Erkennbarkeit der dahinterstehenden Volltexte empirisch wie mathematisch-analytisch zu erforschen.



## Öffentliche Archive und staatliches Wissen

### Die Modernisierung des deutschen Archivrechts

Von Dr. Hannes Berger

2019, 608 S., brosch., 112,- €

ISBN 978-3-8288-4373-8

ePDF 978-3-8288-7355-1

Das deutsche Archivrecht wird reformiert. In der letzten Dekade wird das Ziel verfolgt, die Archivgesetze an die modernen Rahmenbedingungen anzupassen.

Der vorliegende Band untersucht erstmals in systematischer Weise die umfassenden Veränderungen des deutschen Archivwesens zu Beginn des 21. Jahrhunderts. Schwerpunkte sind dabei die Datenschutz-Grundverordnung, die digitale Bestandserhaltung und der Rechtsanspruch der Bürger auf Zugang zu Archivinformationen.

Bestellen Sie im Buchhandel oder  
versandkostenfrei unter **www.tectum-shop.de**

**Tectum  
Verlag**