



Verhältnis von Data Crawling und Urheberrecht im Hochschulkontext

Eine Handreichung der
Rechtsinformationsstelle für die
digitale Lehre bwDigiRecht

18.06.2026

Jana Knecht

Executive Summary

Diese Handreichung erläutert die urheberrechtlichen Rahmenbedingungen für den Einsatz von Data Crawlern, insbesondere im Kontext von Text and Data Mining (TDM) und dem Training von Systemen der Künstlichen Intelligenz (KI) und deren Auswirkungen im Hochschulkontext.¹

► Data Crawling als urheberrechtliche Vervielfältigung

Data Crawling ist ein automatisierter Prozess, bei dem Software-Agenten² Webseiteninhalte auslesen und für eine spätere Verarbeitung technisch fixieren. Urheberrechtlich stellt dieser Vorgang in der Regel eine **Vervielfältigung im Sinne des § 16 des Urhebergesetzes (UrhG)** dar, wobei es unerheblich ist, ob die Kopie nur flüchtig im Arbeitsspeicher oder dauerhaft gespeichert wird.

► Die allgemeine TDM-Schranke (§ 44b UrhG)

Diese Regelung gestattet die automatisierte Analyse von Werken bzw. Text and Data Mining, sofern der Zugang rechtmäßig erfolgt ist und kein wirksamer **Nutzungsvorbehalt** vorliegt. Ein solcher Vorbehalt (Opt-out) muss bei online zugänglichen Inhalten zwingend in einem **maschinenlesbaren Format**, wie etwa in einer robots.txt-Datei, erklärt werden.

¹ Alle hier zitierten Online-Quellen wurden zuletzt am 18.06.2026 abgerufen. Kostenlos abrufbare Medien sind in den Fußnoten und im Literaturverzeichnis verlinkt.

² Ein Software-Agent ist ein Computerprogramm, das autonom im Namen eines Benutzers oder einer Organisation handelt, um bestimmte Aufgaben zu erfüllen. Software-Agenten unterscheiden sich von normalen Computerprogrammen dadurch, dass sie eigenständig handeln und mit anderen Software-Agenten sowie mit Menschen kommunizieren können. Sie können verschiedene Arten von Aufgaben erledigen, wie z.B. die Automatisierung von Prozessen oder die Unterstützung von Benutzern bei der Erledigung ihrer Aufgaben. *Begerow, Software Agents*, [Link](#).

► Das Forschungsprivileg (§ 60d UrhG)

Wissenschaftliche Einrichtungen wie Hochschulen dürfen TDM-Vervielfältigungen für **nicht-kommerzielle Forschung** vornehmen. Da § 60d UrhG zwingendes Recht ist, genießt die Forschung Vorrang vor technischen Sperren; Forschende dürfen maschinenlesbare Nutzungsvorbehalte (Opt-outs) für ihre wissenschaftlichen Analysen daher ignorieren.

► Beweislast und Transparenzpflichten

Im Rahmen der allgemeinen TDM-Schranke liegt die **Beweislast für das Fehlen eines Nutzungsvorbehalts bei den Nutzenden**, was eine lückenlose Dokumentation der Crawling-Vorgänge erforderlich macht. Zudem verpflichtet die **KI-Verordnung** Anbietenden von KI-Modellen zur Veröffentlichung detaillierter Zusammenfassungen ihrer Trainingsdaten, um Rechteinhabenden die Prüfung ihrer Ansprüche zu ermöglichen.

► Abgrenzung zur Lehre (§ 60a UrhG)

Während § 60a UrhG die Nutzung kleiner Werkteile zur Veranschaulichung im Unterricht erlaubt, rechtfertigt diese Norm **nicht das Data Crawling** zum Zweck des KI-Trainings. Für den technischen Vorgang des Crawlens zur Modellentwicklung müssen stets die spezifischen TDM-Schranken gem. §§ 44b, 60d UrhG herangezogen werden.

Inhaltsverzeichnis

01	Data Crawling und urheberrechtliche Einordnung	4
02	Die allgemeine TDM-Schranke (§ 44b UrhG).....	7
2.1	Voraussetzungen und Anwendungsbereich.....	7
2.2	Nutzungsvorbehalt	8
03	Das Forschungsprivileg gemäß § 60d UrhG	11
04	Lizenzfragen	13
4.1	Open-Source.....	13
4.2	Creative Commons (CC-Lizenzen)	14
4.3	Individuelle Lizenzbedingungen (AGB)	15
05	Beweislast und Haftungsrisiken	17
5.1	Beweislastverteilung nach § 44b Abs. 3 UrhG	17
5.2	Der Indizienbeweis durch KI-Output.....	18
5.3	Einfluss der KI-Verordnung auf die Beweissituation	18
5.4	Haftung für Output-Verletzungen	18
5.5	Besonderheiten im Forschungskontext an Hochschulen	19
06	Abgrenzung zur Lehre gemäß § 60a UrhG	20
07	Fazit und Ausblick	22
08	Literaturverzeichnis.....	24

01 Data Crawling und urheberrechtliche Einordnung

Data Crawling, oft auch als Web-Crawling oder Indexierung bezeichnet,³ ist der automatisierte Prozess, bei dem spezialisierte Software-Agenten⁴, sogenannte Bots oder Crawler, Webseiten im Internet systematisch besuchen, die Struktur der Webseiten analysieren und die dort befindlichen Inhalte auslesen.⁵ Technisch gesehen navigieren diese Programme von Link zu Link, um eine Kopie der besuchten Inhalte zu erstellen und diese für eine spätere Verarbeitung verfügbar zu machen.⁶

Dieser Vorgang geht über das bloße Betrachten einer Webseite durch einen menschliche Nutzende hinaus, da die Daten für eine automatisierte Analyse technisch extrahiert und in einem Speicher fixiert werden müssen. Dadurch

³ Zu unterscheiden von Web Scraping: Während beim Web Crawling automatisiert Webseiten systematisch durchsucht und über Links neue Seiten entdeckt werden (Navigation von Link zu Link), nutzt Web Scraping bereits bekannte bzw. gesammelte URLs, um gezielt bestimmte Inhalte oder Daten aus den Seiten zu extrahieren. Web Crawling dient somit primär der Erkundung und Indexierung des Webs, während Web Scraping auf die inhaltliche Datengewinnung aus einzelnen Seiten abzielt. Mehr dazu: *Kadek*, Web Scraping vs. Web Crawling: Was ist der Unterschied? Ein umfassender Vergleich, [Link](#).

⁴ Ein Software-Agent ist ein Computerprogramm, das autonom im Namen eines Benutzers oder einer Organisation handelt, um bestimmte Aufgaben zu erfüllen. Software-Agenten unterscheiden sich von normalen Computerprogrammen dadurch, dass sie eigenständig handeln und mit anderen Software-Agenten sowie mit Menschen kommunizieren können. Sie sind in der Lage, selbstständig Aktionen zu durchführen, um ihre Ziele zu erreichen, und können auch lernen, indem sie auf zuvor getätigte Entscheidungen und Beobachtungen basieren. Begerow, Software Agents, [Link](#).

⁵ Datenschutz im Internet, I. Grundlagen des Internets, Rn. 32-34; zu Indexierung Fn. 40: *Witte Paz/Seitz*, NZKart 2025, [Link](#).

⁶ Datenschutz im Internet, I. Grundlagen des Internets, 3. Webdienste, Rn. 33.

kann eine Art von temporärer oder dauerhafter digitaler Datenbank entstehen.

Urheberrechtlich stellt dieser Vorgang in der Regel eine **Vervielfältigung im Sinne des § 16 UrhG** dar.⁷ Eine solche Vervielfältigung liegt bei jeder körperlichen Festlegung eines Werkes vor, die geeignet ist, das Werk den menschlichen Sinnen auf irgendeine Art mittelbar oder unmittelbar wahrnehmbar zu machen.⁸ Da Crawler die Inhalte zum Zweck der Auswertung technisch fixieren, sei es flüchtig im Arbeitsspeicher oder dauerhaft in einem KI-Trainingskorpus, wird der Tatbestand der Vervielfältigung erfüllt.

Die Dauer der Vervielfältigung ist dabei unerheblich, da sowohl vorübergehende als auch dauerhafte Kopien vom Vervielfältigungsrecht der Urhebenden erfasst werden.⁹

Technisch bildet das Data Crawling die notwendige Vorstufe für das **Text and Data Mining (TDM)**. TDM wird in §44b UrhG definiert als die automatisierte Analyse von einzelnen oder mehreren digitalen bzw. digitalisierten Werken, um insbesondere Informationen über **Muster, Trends und Korrelationen** zu gewinnen. Dieser Analyseprozess ist von zentraler Bedeutung für die Entwicklung und das Training von KI, da KI-Modelle auf der statistischen Auswertung sehr großen Datenmengen basieren, die zuvor mittels Crawling gesammelt wurden.¹⁰

Ob die beim Crawling stattfindenden Vervielfältigungen zulässig sind, richtet sich im deutschen Recht primär nach den Schrankenregelungen der **§§ 44b und 60d UrhG**.¹¹ Während § 44b UrhG eine **allgemeine Erlaubnis** für TDM-Vervielfältigungen vorsieht, die durch einen maschinenlesbaren Nutzungsvorbehalt ausgeschlossen werden kann, privilegiert § 60d UrhG

⁷ Maamar, Z. Für Urh.- Medien. 2023, 481 (484).

⁸ LG München, GRUR 2025, [Link](#), Rn. 180.

⁹ LG München, GRUR 2025, [Link](#), Rn. 179.

¹⁰ Grimm/Münster, GRUR-Prax 2025, (443), [Link](#), Rn. 1-3.

¹¹ Baumann, Neue Juristische Wochenschr. 2023, 3673 (3673), Rn. 3.

die **wissenschaftliche Forschung** und gestattet TDM-Handlungen selbst gegen den ausdrücklichen Willen der Rechteinhabenden.

Die Abgrenzung zwischen diesen urheberrechtlichen Normen ist im Hochschulkontext essenziell, um die rechtssichere Nutzung von Data Crawlern beispielsweise für Forschungsprojekte zu gewährleisten.

02 Die allgemeine TDM-Schranke (§ 44b UrhG)

Die allgemeine Schrankenregelung des **§ 44b UrhG** gestattet TDM-Vervielfältigungen grundsätzlich für jedermann, sofern der Zugang zum Werk rechtmäßig erfolgt ist und kein wirksamer **Nutzungsvorbehalt** vorliegt.¹² Ein solcher Vorbehalt muss bei onlinezugänglichen Inhalten in einem **maschinenlesbaren Format** erklärt werden, um gegenüber automatisierten Crawlern rechtlich verbindlich zu sein.¹³ Sobald ein solcher Vorbehalt wirksam erklärt wurde, entfällt die gesetzliche Erlaubnis zur Nutzung, wodurch die Vervielfältigung wieder lizenzpflichtig wird.¹⁴

2.1 Voraussetzungen und Anwendungsbereich

Die Zulässigkeit von Vervielfältigungen nach § 44b UrhG setzt voraus, dass die Nutzenden einen **rechtmäßigen Zugang** zu dem Werk haben.¹⁵ Ein rechtmäßiger Zugang ist grundsätzlich dann gegeben, wenn die Nutzenden die Möglichkeit haben, das Werk in einer Weise wahrzunehmen, die im Einklang mit dem Gesetz steht. Dies umfasst beispielsweise das Betrachten einer Webseite im Browser, das Ausleihen eines Buches in einer Bibliothek oder den Zugriff auf Inhalte im Rahmen eines bestehenden Abonnements.¹⁶

Ein rechtmäßiger Zugang ist im Internetkontext weit auszulegen; er liegt bereits dann vor, wenn Inhalte ohne technische Schutzmaßnahmen oder Bezahlschranken frei abrufbar sind. Solange Rechteinhabende Werke nicht

¹² Käde, Künstl. Intell. Recht 2024, 162 (163 f.).

¹³ Bomhard, Urheberrecht, UrhG § 44b, Rn. 25 f., Rn. 32 ff.

¹⁴ Baumann, Neue Juristische Wochenschr. 2023, 3673 (3675).

¹⁵ Pesch, Recht der elektronischen Medien, UrhG § 44b, Rn. 27 ff.

¹⁶ Anton, Recht der elektronischen Medien, UrhG § 60d, Rn. 30-32.

hinter eine Login- oder Paywall-Schranke stellen, darf auf die Werke zu TDM-Zwecken zugegriffen werden.¹⁷ Die Grenze der rechtmäßigen Zugänglichkeit wird dort gezogen, wo die Quelle des Werkes **offensichtlich rechtswidrig** ist. Ein Beispiel hierfür sind illegale Filesharing-Plattformen, bei denen der Zugriff nicht auf einer gesetzlichen Erlaubnis oder einer Zustimmung der Urhebenden beruht; solche Quellen vermitteln keinen rechtmäßigen Zugang im Sinne der TDM-Schranken.¹⁸

Die entscheidende Einschränkung der allgemeinen TDM-Schranke findet sich in **§ 44b Abs. 3 UrhG**: Die Nutzung ist unzulässig, wenn Rechteinhaber sich diese ausdrücklich vorbehalten haben.¹⁹ Dieser Vorbehalt ermöglicht es Urhebenden und Webseitenbetreibenden, das automatisierte Auslesen ihrer Inhalte durch Data Crawler zu untersagen.²⁰

2.2 Nutzungsvorbehalt

Damit ein Nutzungsvorbehalt gegenüber Data Crawlern wirksam ist, muss er bei onlinezugänglichen Inhalten zwingend in einem **maschinenlesbaren Format** erfolgen. Die Zulässigkeit der Vervielfältigung entfällt jedoch, wenn der Rechteinhaber einen **Nutzungsvorbehalt** erklärt hat.²¹ Dieser Vorbehalt muss ausdrücklich erfolgen. Bei Werken, die online zugänglich sind, ist ein solcher Vorbehalt nur dann wirksam, wenn er in einem **maschinenlesbaren Format** erfolgt. Hierfür haben sich technische Standards wie die Hinterlegung in der **robots.txt-Datei** oder die Verwendung von Metadaten im HTML-Code etabliert. Widersprüchliche Willenserklärungen, die lediglich in natürlicher Sprache auf einer Webseite verfasst sind, bleiben gegenüber automatisierten Crawlern im Rahmen des § 44b UrhG rechtlich unbeachtlich.²²

In diesem Kontext sind die folgenden technischen Konzepte maßgeblich:

¹⁷ Pesch, Recht der elektronischen Medien, UrhG § 44b, Rn. 30 ff.

¹⁸ Käde, DSRITB 2024, (8), [Link](#).

¹⁹ Pesch, Recht der elektronischen Medien, UrhG § 44b, Rn. 39.

²⁰ Antoine, GRUR 2025, (123), [Link](#).

²¹ Käde, DSRITB 2024, (12), [Link](#).

²² Götting u. a., Urheberrecht, [Link](#), UrhG § 44b, Rn. 32 ff.

- **Technische Standards:** Ein Nutzungsvorbehalt gilt als maschinenlesbar, wenn seine Inhalte durch Softwareprogramme, die für das Data Mining eingesetzt werden, automatisiert ausgelesen und identifiziert werden können.²³
- **robots.txt:** In der gängigen Praxis wird hierfür die robots.txt-Datei verwendet, die im Hauptverzeichnis einer Webseite hinterlegt ist und Anweisungen an automatisierte Bots enthält. Bei einer solchen Datei handelt es sich um eine schlichte Textdatei, die im Hauptverzeichnis einer Webseite hinterlegt wird, um mit automatisierten Programmen zu kommunizieren.²⁴ Diese Datei dient dazu, Crawlern oder Bots mitzuteilen, ob und welche Bereiche einer Webseite für eine automatisierte Indexierung oder Analyse zugänglich sind. Technisch funktioniert dies so, dass ein Crawler vor dem Auslesen der eigentlichen Inhalte die robots.txt ansteuert und die dort hinterlegten Anweisungen identifiziert.²⁵ Durch die Adressierung spezifischer Bot-Namen können die Betreibenden so gezielte Verbote für bestimmte Crawler aussprechen, was im Rahmen des § 44b Abs. 3 UrhG die gängige Methode, deren Richtigkeit noch zu prüfen ist,²⁶ zur Erklärung eines maschinenlesbaren Nutzungsvorbehalts darstellt.²⁷
- **Ausschluss von natürlicher Sprache:** Ein bloßer urheberrechtlicher Hinweis in natürlicher Sprache auf der Webseite oder in den Allgemeinen Geschäftsbedingungen (AGB) genügt den Anforderungen an die Maschinenlesbarkeit nicht und ist gegenüber automatisierten Crawlern im Rahmen des § 44b UrhG unbeachtlich.²⁸
- **TDM Reservation Protocol:** Ein zentraler Ansatz zur Standardisierung ist das **TDM Reservation Protocol**, welches von einer Community-Gruppe des

²³ Pesch, Recht der elektronischen Medien, UrhG § 44b, Rn. 40.

²⁴ Götting u. a., Urheberrecht, [Link](#), UrhG § 44b, Rn. 32-34.

²⁵ Schmidt/Pruß, Handbuch IT- und Datenschutzrecht, § 3 Technische Grundlagen des Internets, RN. 403-419.

²⁶ Bomhard, Urheberrecht, § 44b, Rn. 32-34.

²⁷ Käde, DSRITB 2024, (6), [Link](#).

²⁸ Maamar, Z. Für Urh.- Medien. 2023, 481 (484).

World Wide Web Consortium (W3C) entwickelt wurde.²⁹

Dieses Protokoll ermöglicht es Rechteinhabenden, einen Nutzungsvorbehalt durch eine binäre Kodierung (0 oder 1) direkt im HTML-Quellcode einer Webseite oder in den HTTP-Headern zu hinterlegen.³⁰ Der Rückgriff auf einen **W3C-Standard** dient der Herstellung von Rechtssicherheit für KI-Anbietende und Webseitenbetreibende gleichermaßen.³¹ Da Data Crawler darauf angewiesen sind, Verbote ohne menschliche Interpretation zu erkennen, stellen solche Protokolle sicher, dass ein Nutzungsvorbehalt angemessen im Sinne des Art. 4 Abs. 3 der europäischen Richtlinie über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt (DSM-Richtlinie)³² und des § 44b UrhG erklärt wurde.³³

Für **Hochschulen und Forschungseinrichtungen** ist zu beachten, dass diese technischen Protokolle im Rahmen des Forschungsprivilegs nach **§ 60d UrhG** zwar ausgelesen, aber für die Analysephase bzw. Input ignoriert werden dürfen, da § 60d UrhG als zwingendes Recht den Vorrang der Wissenschaft gegenüber solchen technischen Sperren sichert.³⁴ Dennoch bleibt die Kenntnis dieser Standards für die Compliance bei kommerziellen Drittmittelprojekten oder der späteren Veröffentlichung von Datensätzen unerlässlich.

²⁹ *World Wide Web Consortium (W3C)*, TDM Reservation Protocol (TDMRep), [Link](#).

³⁰ *Maamar*, Z. Für Urh.- Medien. 2023, 481 (484).

³¹ Vgl. *Ammann u. a.*, KIR 2026, (13), [Link](#).

³² *Europäisches Parlament und Rat der Europäischen Union*, Verordnung (EU) 2022/2065 des Europäischen Parlaments und des Rates vom 19. Oktober 2022 über einen Binnenmarkt für digitale Dienste und zur Änderung der Richtlinie 2000/31/EG (Gesetz über digitale Dienste), [Link](#).

³³ *Maamar*, Z. Für Urh.- Medien. 2023, 481 (485).

³⁴ *Anton*, Recht der elektronischen Medien, UrhG § 60d, Rn. 35; *Maamar*, Z. Für Urh.- Medien. 2023, 481 (485).

03 Das

Forschungsprivileg gemäß § 60d UrhG

Die Regelung des **§ 60d UrhG** bildet das urheberrechtliche Fundament für datengestützte Wissenschaft an Hochschulen, indem sie Vervielfältigungen für Text and Data Mining zu Zwecken der **wissenschaftlichen Forschung** ausdrücklich privilegiert. Diese Bestimmung dient der Umsetzung von Art. 3 DSM-Richtlinie und soll sicherstellen, dass die automatisierte Auswertung großer Datenmengen für die Wissensgenerierung nicht durch urheberrechtliche Exklusivrechte blockiert wird.³⁵

Das Privileg des § 60d UrhG steht nicht jedermann offen, sondern ist an den Status der nutzenden Institution gebunden. **Forschungsorganisationen**, worunter insbesondere Hochschulen, öffentliche Bibliotheken und andere Einrichtungen fallen, die wissenschaftliche Forschung betreiben, dürfen diese Schranke beanspruchen.³⁶ Voraussetzung ist dabei stets, dass die Forschung zu **nicht-kommerziellen Zwecken** erfolgt oder die Einrichtung im öffentlichen Interesse handelt.³⁷ Sofern Forschungsprojekte in enger Kooperation mit privaten Unternehmen durchgeführt werden, muss im Einzelfall geprüft werden, ob der nicht-kommerzielle Charakter der wissenschaftlichen Tätigkeit noch überwiegt, um die Privilegierung nicht zu verlieren.³⁸

³⁵ Urheberrechtsgesetz, [Link](#), 60d, Rn. 1 f.

³⁶ *Maamar*, Z. Für Urh.- Medien. 2023, 481 (485).

³⁷ Urheberrechtsgesetz, [Link](#), UrhG § 60d, Rn. 2, 4-7.

³⁸ Vgl. Urheberrechtsgesetz, [Link](#), UrhG § 60d, Rn. 5.

Ein wesentlicher Unterschied zur allgemeinen TDM-Schranke des § 44b UrhG besteht darin, dass § 60d UrhG **zwingendes Recht** darstellt und nicht durch einen Nutzungsvorbehalt der Rechteinhabenden ausgeschlossen werden kann.³⁹ Selbst, wenn die Webseitenbetreibenden technische Sperren in der **robots.txt-Datei** implementieren oder ausdrückliche Opt-out-Erklärungen in maschinenlesbarer Form⁴⁰ hinterlegt haben, dürfen Forschungseinrichtungen diese für ihre wissenschaftlichen Analysen ignorieren. Dieser Vorrang der Forschung gilt jedoch nur für den Vorgang der Analyse (Input) und die Erstellung des Korpus; er entbindet nicht von der Pflicht, den rechtmäßigen Zugang (**siehe Abschnitt 2.1.**) zur Quelle sicherzustellen.⁴¹

In der Praxis bedeutet dies, dass Hochschulangehörige Data Crawler einsetzen dürfen, um urheberrechtlich geschützte Daten systematisch zu sammeln und damit eigene KI-Modelle zu Trainingszwecken zu füttern, sofern dies einen wissenschaftlichen Forschungszweck verfolgt. Sobald die so entwickelten Modelle jedoch kommerziell verwertet werden sollen, verlässt die Nutzung den Schutzbereich des § 60d UrhG, und es müssen entweder Lizenzen eingeholt oder die strengeren Anforderungen des § 44b UrhG, einschließlich der Beachtung von Opt-outs, erfüllt werden.⁴²

³⁹ *Maamar*, Z. Für Urh.- Medien. 2023, 481 (486).

⁴⁰ Mehr zu der „maschinenlesbar Form“ *Knecht*, Rechtsbereichsspezifische Betrachtung von KI: Urheberrecht, S. 7, [Link](#).

⁴¹ *Maamar*, Z. Für Urh.- Medien. 2023, 481 (485).

⁴² *Wendt/Wendt*, Das neue Recht der Künstlichen Intelligenz, § 11 KI mit allgemeinem Verwendungszweck - GPAI, Rn. 40 ff.

04 Lizenzfragen

Das Verhältnis zwischen vertraglichen **Lizenzfragen** und den gesetzlichen Schrankenregelungen zum **Text and Data Mining** bildet einen zentralen Streitpunkt in der aktuellen urheberrechtlichen Debatte. Da dieser Prozess technisch bedingt zwingend Vervielfältigungen gemäß § 16 UrhG erfordert, stellt sich die Frage, ob Nutzende hierfür eine individuelle Lizenz erwerben müssen oder ob die gesetzlichen Erlaubnistatbestände eingreifen.⁴³

Rechtlich ist dabei zu beachten, dass das bloße Auffinden frei zugänglicher Inhalte nicht zwingend deren rechtmäßige Nutzung im Sinne des Urheberrechts oder einschlägiger Lizenzbedingungen erlaubt. Vielmehr ist im Einzelfall zu prüfen, ob und in welchem Umfang die erfassten Daten tatsächlich unter offenen Lizenzen stehen oder ob Nutzungseinschränkungen, z.B. durch spezifische Lizenzklauseln oder technische Schutzmaßnahmen, bestehen. Der Einsatz von Data-Crawlern stellt zwar ein legitimes technisches Mittel dar, muss jedoch stets von einer sorgfältigen Lizenz- und Rechteprüfung begleitet werden muss, um Rechtsverletzungen zu vermeiden.

4.1 Open-Source

Open-Source-Lizenzen basieren auf dem Urheberrecht, gewähren den Nutzenden jedoch weitreichende Freiheiten zur Nutzung, Bearbeitung und Verbreitung des Werkes, oft unter der Bedingung, dass Änderungen unter denselben Bedingungen geteilt werden.⁴⁴

⁴³ Leistner, GRUR 2025, (1129), [Link](#); LG München, GRUR 2025, [Link](#), Rn. 100, 109, 111.

⁴⁴ Urheberrechtsgesetz, [Link](#), UrhG § 69c, Rn. 38.

In der KI-Entwicklung gewinnen diese Lizenzen insbesondere im Kontext von **Open Weights**⁴⁵ an Bedeutung, bei denen die trainierten Parameter eines Modells unter einer freien Lizenz veröffentlicht werden.⁴⁶ Die **KI-Verordnung** sieht für solche Open-Source-Modelle Erleichterungen bei den Transparenzpflichten vor, sofern sie keine systemischen Risiken bergen. Dennoch entbindet die Nutzung von Open-Source-Daten den Anbietenden nicht von der Pflicht, eine Strategie zur Einhaltung des Urheberrechts zu implementieren, was insbesondere die Identifizierung von Nutzungsvorbehalten Dritter im Trainingsdatensatz einschließt.⁴⁷

4.2 Creative Commons (CC-Lizenzen)

Creative-Commons-Lizenzen sind standardisierte Lizenzverträge, die es Urhebenden ermöglicht, der Allgemeinheit auf einfache Weise Nutzungsrechte einzuräumen.⁴⁸ Sie bestehen aus modularen Bausteinen, deren Bedeutung für das KI-Training unterschiedlich bewertet wird:

- **CC0 (Public Domain Dedication):** Das Werk wird in die Gemeinfreiheit entlassen.⁴⁹ Data Crawler können diese Inhalte ohne rechtliche Hürden für Training und Analyse einlesen.⁵⁰

⁴⁵ Vgl. zur urheberrechtlichen Einordnung sogenannter „Open Weights“-Modelle, bei denen die trainierten Modellparameter öffentlich zugänglich gemacht werden, aber Trainingsdaten, Trainingsprozesse oder Quellcode nicht notwendigerweise vollständig offengelegt sind, die Diskussion um den Werk- bzw. Schutzgegenstandscharakter von KI-Modellen sowie deren mögliche Einordnung als Computerprogramme i.S.d. §§ 69a ff. UrhG oder als Datenbanken i.S.d. §§ 87a ff. UrhG. Problematisch ist insbesondere, inwieweit die Gewichte selbst eine hinreichende Schöpfungshöhe aufweisen oder lediglich als Ergebnis eines (teil-)automatisierten Prozesses zu qualifizieren sind. Ebenso umstritten ist, ob und in welchem Umfang urheberrechtlich geschützte Inhalte aus den Trainingsdaten in den Modellparametern fortwirken („output memorization“) und dadurch Vervielfältigungstatbestände berührt werden. Ziegler/Jürges, RD 2026 77 Ff 2026, 77, Rn. 3 f.

⁴⁶ Ziegler/Jürges, RD 2026 77 Ff 2026, 77, Rn. 3.

⁴⁷ Ziegler/Jürges, RD 2026 77 Ff 2026, 77, Rn. 6-9.

⁴⁸ Urheberrecht Fromm/Nordemann, [Link](#), Vorbemerkung §§ 31 ff., k] Open Content, Rn. 330b.

⁴⁹ *Creative commons*, About CC Licences, [Link](#), über CC0 1.0 Universell.

⁵⁰ Vgl. Urheberrecht Fromm/Nordemann, [Link](#), Vorbemerkung §§ 31 ff., k] Open Content, Rn. 330b.

- **NC (Non-Commercial)⁵¹ und ND (No Derivatives)⁵²:**

Die Nutzung von NC-lizenzierten Inhalten für das Training kommerzieller Modelle ist umstritten, da diese Bedingungen als vertragliche Nutzungsvorbehalte gewertet werden könnten.⁵³ Während die Forschung an Hochschulen durch § 60d UrhG privilegiert bleibt, stellt die kommerzielle Verwertung solcher Daten einen potenziellen Lizenzverstoß dar.⁵⁴ Zudem ist die Einhaltung der **Attributionsspflicht (BY)⁵⁵** bei Milliarden von Trainingsdaten technisch kaum realisierbar, falls Teile des geschützten Werkes im KI-Output erkennbar bleiben.⁵⁶

4.3 Individuelle Lizenzbedingungen (AGB)

Individuelle Lizenzbedingungen sind rein vertragliche Vereinbarungen zwischen den Anbietenden eines Dienstes und dessen Nutzenden, die oft in Form von **AGB** vorliegen.

Charakteristisch ist hier die **Relativität der Schuldverhältnisse**: Diese Bedingungen binden rechtlich nur die Vertragsparteien. Verpflichten sich die Nutzenden beispielsweise gegenüber OpenAI, nur Daten einzugeben, an denen er alle Rechte hält, begründet ein Verstoß hiergegen eine Vertragsverletzung gegenüber OpenAI. Demnach käme OpenAI ein vertraglicher Schadensersatzanspruch zu, den die eigentlichen Urheberrechtsinhabenden mangels Vertrags mit den Nutzenden nicht geltend machen kann. Die Urheberrechtsinhabenden haben dennoch gesetzliche Schadensersatzansprüche.⁵⁷

⁵¹ *Creative commons*, About CC Licences, [Link](#), über Attribution-NonCommercial 4.0 International.

⁵² *Creative commons*, About CC Licences, [Link](#), über Attribution-NoDerivatives 4.0 International.

⁵³ *Schippel*, Künstl. Intell. Recht 2026, 36 (38).

⁵⁴ *Morgenroth*, Hochschulstudienrecht und Hochschulprüfungsrecht, B. Hochschulstudienrecht, 2 dd], Rn. 377.

⁵⁵ *Creative commons*, About CC Licences, [Link](#), über Attribution 4.0 International.

⁵⁶ *Schippel*, Künstl. Intell. Recht 2026, 36 (38).

⁵⁷ *Maamar*, Z. Für Urh.- Medien. 2023, 481 (491).

Für Hochschulen ist entscheidend, dass Mitarbeitende bevorzugt institutionell bereitgestellte KI-Systeme nutzen sollten, da bei der Verwendung privater Accounts für dienstliche Zwecke das Risiko einer **persönlichen Haftung** bei Verstößen gegen die AGB der Anbietenden besteht.⁵⁸ Zudem verpflichtet die KI-Verordnung Anbietenden von GPAI-Modellen zur Veröffentlichung einer detaillierten Zusammenfassung der Trainingsdaten, um Rechteinhabenden die Prüfung ihrer vertraglichen und gesetzlichen Ansprüche überhaupt erst zu ermöglichen.⁵⁹

⁵⁸ Schlotfeldt, 10 Fragen zu KI-Output, Urheberrecht und OER, S. 33, [Link](#).

⁵⁹ Wendt/Wendt, Das neue Recht der Künstlichen Intelligenz, § 11 KI mit allgemeinem Verwendungszweck - GPAI, Rn. 49.

05 Beweislast und Haftungsrisiken

Die prozessuale Durchsetzung von Urheberrechtsansprüchen im Zusammenhang mit Data Crawling und KI-Training unterliegt spezifischen Beweislastregeln, die für die Beteiligten erhebliche Herausforderungen mit sich bringen. Diese Risiken betreffen sowohl die Phase der Datenbeschaffung als auch Output, also die Generierung von Inhalten durch das KI-System.

5.1 Beweislastverteilung nach § 44b Abs. 3 UrhG

Im Rahmen der allgemeinen TDM-Schranke des § 44b UrhG liegt die **Beweislast für das Fehlen eines Nutzungsvorbehalts bei den Nutzenden** des Werkes.⁶⁰ Werden urheberrechtlich geschützte Daten durch Crawler für das Training von KI-Systemen vervielfältigt, müssen die Betreibenden im Streitfall nachweisen, dass zum Zeitpunkt der Nutzung kein wirksamer, maschinenlesbarer Vorbehalt vorlag.⁶¹ In der rechtswissenschaftlichen Literatur wird betont, dass dieser Beweis angesichts der enormen Datenmengen, die oft Milliarden von Einzelwerken umfassen, für jeden einzelnen Inhalt **praktisch kaum zu führen ist**.⁶² Daher wird KI-Anbietenden und Forschungseinrichtungen geraten, ihre technischen Sicherungsmechanismen zur Erkennung von Opt-outs umfassend zu dokumentieren, um ihre Sorgfalt nachweisen zu können.⁶³

⁶⁰ Maamar, Z. Für Urh.- Medien. 2023, 481 (485).

⁶¹ Grimm/Münster, GRUR-Prax 2025, (445), [Link](#), Rn. 27.

⁶² Vgl. Gerecke, GRUR-Prax 2023, (382 f.), [Link](#).

⁶³ Maamar, Z. Für Urh.- Medien. 2023, 481 (485).

5.2 Der Indizienbeweis durch KI-Output

Ein zentrales Haftungsrisiko ergibt sich daraus, dass der Output eines KI-Modells Rückschlüsse auf die verwendeten Trainingsdaten zulässt.⁶⁴ Wenn ein KI-System Ergebnisse generiert, die eine **hohe Ähnlichkeit zu einem geschützten Werk** aufweisen, gilt dies nach aktueller Ansicht als ein **entscheidendes Indiz** dafür, dass dieses Werk tatsächlich für das Training verwendet und damit vervielfältigt wurde.⁶⁵ Die Wiedererkennbarkeit ermöglicht es Rechteinhabenden, eine Urheberrechtsverletzung auf der Input-Seite substantiiert zu behaupten, ohne direkten Zugriff auf den Trainingskorpus zu haben.⁶⁶

5.3 Einfluss der KI-Verordnung auf die Beweissituation

Die **Transparenzpflichten** der KI-Verordnung flankieren diese Beweissituation, indem sie Anbietenden von GPAI-Modellen zur Veröffentlichung einer hinreichend detaillierten Zusammenfassung ihrer Trainingsdaten verpflichten. Dies soll die bestehende Informationsasymmetrie zwischen Urhebenden und KI-Entwickelnden verringern und es Rechteinhabenden ermöglichen, unbefugte Nutzungen ihrer Werke durch Data Crawler effektiver nachzuvollziehen.⁶⁷ Die Nichterfüllung dieser Dokumentationspflichten kann dabei regulatorische Sanktionen durch das **EU-KI-Büro** nach sich ziehen und die prozessuale Position der Anbietenden schwächen.⁶⁸

5.4 Haftung für Output-Verletzungen

Neben der Vervielfältigung während des Trainings besteht das Risiko einer Haftung für den **KI-generierten Output** selbst.⁶⁹ Sofern ein KI-Output

⁶⁴ Gerecke, GRUR-Prax 2023, (383 f.), [Link](#).

⁶⁵ LG München, GRUR 2025, [Link](#), Rn. 63, 263.

⁶⁶ Vgl. Maamar, Z. Für Urh.- Medien. 2023, 481 (489 f.).

⁶⁷ Ammann u. a., KIR 2026, (13), [Link](#).

⁶⁸ Antoine, GRUR 2025, (127), [Link](#).

⁶⁹ LG München, GRUR 2025, [Link](#), Rn. 220.

wesentliche eigenpersönliche Züge eines geschützten Werkes übernimmt, kann dies eine Urheberrechtsverletzung durch den Nutzenden oder den Anbietenden des KI-Systems darstellen.⁷⁰ Ob eine Haftung nach den Grundsätzen der Störerhaftung eintritt, wird im Einzelfall anhand der Beherrschbarkeit des Generierungsprozesses beurteilt.⁷¹

5.5 Besonderheiten im Forschungskontext an Hochschulen

Hochschulen, die Data Crawling im Rahmen des **Forschungsprivilegs** gem. **§ 60d UrhG** betreiben, tragen die Beweislast für die Einhaltung der Privilegierungsvoraussetzungen.⁷² Sie müssen beweisen können, dass der Zugang zu den Quellen rechtmäßig war und die Analyse ausschließlich **nicht-kommerziellen wissenschaftlichen Zwecken** dient.⁷³ Da § 60d UrhG gegenüber Nutzungsvorbehalten vorrangig ist, entfällt im reinen Forschungskontext die Beweislast hinsichtlich der Beachtung technischer Sperren wie robots.txt.⁷⁴

⁷⁰ Wagner, Z. Für Recht Digit. Datenwirtschaft IT, 298 (300).

⁷¹ Leistner, GRUR 2025, (1133), [Link](#).

⁷² Vgl. Anton, Recht der elektronischen Medien, UrhG § 60d, Rn. 19.

⁷³ Bomhard, Urheberrecht, UrhG § 60d, Rn. 14; Anton, Recht der elektronischen Medien, UrhG § 60d, Rn. 28.

⁷⁴ Vgl. LG München, GRUR 2025, [Link](#), Rn. 99.

06 Abgrenzung zur Lehre gemäß § 60a UrhG

Während die Schranken des Text and Data Mining aus §§ 44b, 60d UrhG primär den technischen Prozess der Datenanalyse und des KI-Trainings auf Input-Ebene adressieren, regelt **§ 60a UrhG** die zulässige Nutzung urheberrechtlich geschützter Inhalte zur **Veranschaulichung von Unterricht und Lehre**. Diese Bestimmung erlaubt es Lehrenden an Bildungseinrichtungen, bis zu **15 % eines veröffentlichten Werkes** für einen abgegrenzten Teilnehmerkreis, z. B. Studierende einer spezifischen Vorlesung, zu vervielfältigen, zu verbreiten und öffentlich zugänglich zu machen.⁷⁵

Die Abgrenzung ist insbesondere dann relevant, wenn der Output eines KI-Systems geschützte Fragmente von Werken Dritter enthält oder diesen in einer Weise ähnelt, die eine **Wiedererkennbarkeit** begründet.⁷⁶ In solchen Fällen kann die Veröffentlichung des KI-Erzeugnisses im Internet eine Urheberrechtsverletzung darstellen. Eine Nutzung innerhalb der Hochschule kann jedoch durch § 60a UrhG privilegiert sein, sofern sie der Veranschaulichung des Lehrstoffs dient.⁷⁷

Für die Praxis an Hochschulen ergeben sich aus dieser Schranke spezifische Anforderungen:

⁷⁵ Anton, Recht der elektronischen Medien, § 60a, Rn. 26 ff.

⁷⁶ Wagner, Z. Für Recht Digit. Datenwirtschaft IT, 298 (300 f.).

⁷⁷ Urheberrecht Fromm/Nordemann, [Link](#), UrhG § 60a, Rn. 4.

- **Beschränkter Teilnehmendenkreis:** Die Bereitstellung solcher Materialien darf ausschließlich für die Teilnehmenden der jeweiligen Lehrveranstaltung erfolgen.⁷⁸
- **Nutzungsort:** Nutzende müssen sich bei grenzüberschreitenden digitalen Lehrangeboten nur an das Recht des Sitzstaates ihrer Bildungseinrichtung halten.⁷⁹ Dies ist besonders relevant für das **Distance Learning**, da die Erlaubnis unabhängig davon gilt, wo sich Lehrende oder Lernende physisch aufhalten.⁸⁰
- **Passwortschutz:** Technisch muss der Zugang durch wirksame Schutzmaßnahmen, z. B. innerhalb eines Lernmanagementsystems wie Moodle oder ILIAS, auf diesen Kreis begrenzt werden.⁸¹
- **Quellenangabe:** Auch bei der Nutzung nach § 60a UrhG bleibt die Pflicht zur Urheber- und Quellenangabe bestehen,⁸² was bei KI-generierten Inhalten, deren Trainingsgrundlage oft unklar ist, eine besondere Herausforderung darstellen kann.

Häufig wird gefragt, ob Data Crawling für KI-Modelle durch § 60a UrhG gerechtfertigt sein kann. Dies ist zu verneinen. § 60a UrhG privilegiert lediglich die Nutzung zur Veranschaulichung im Unterricht für einen begrenzten Teilnehmendenkreis. Der technische Vorgang des Crawlens zum Zweck des KI-Trainings dient jedoch nicht unmittelbar dieser Veranschaulichung, weshalb hierfür ausschließlich die TDM-Schranken gem. §§ 44b, 60d UrhG heranzuziehen sind.

⁷⁸ Urheberrecht Fromm/Nordemann, [Link](#), UrhG § 60a, Rn. 7.

⁷⁹ Anton, Recht der elektronischen Medien, UrhG § 60a, Rn. 43.

⁸⁰ Urheberrecht Fromm/Nordemann, [Link](#), § 60a, Rn. 9b.

⁸¹ Schlotfeldt, 10 Fragen zu KI-Output, Urheberrecht und OER, S. 12, [Link](#), Frage 02.

⁸² Urheberrecht Fromm/Nordemann, [Link](#), UrhG § 60a, Rn. 11.

07 Fazit und Ausblick

Zusammenfassend lässt sich festhalten, dass das Verhältnis von Data Crawlern und Urheberrecht im Hochschulkontext von einem Spannungsverhältnis zwischen dem Schutz geistigen Eigentums und der Freiheit der Forschung geprägt ist. Während das automatisierte Auslesen von Daten technisch die Grundlage für moderne Forschung und KI-Entwicklung bildet, stellt es rechtlich einen massiven Eingriff in das Vervielfältigungsrecht der Urhebenden dar.

Für Hochschulen bietet das Forschungsprivileg des § 60d UrhG jedoch einen robusten Rechtsrahmen, der weitgehende Freiheiten für die nicht-kommerzielle wissenschaftliche Analyse ermöglicht. Der Gesetzgeber hat hier bewusst entschieden, dass wissenschaftlicher Erkenntnisgewinn Vorrang vor individuellen Nutzungsvorbehalten hat, sofern der Zugang zu den Quellen rechtmäßig erfolgt ist. Damit bleibt die Wissenschaft ein geschützter Raum für datenintensive Innovationen.

Gleichzeitig verschärfen sich die Anforderungen an die Transparenz und Compliance, insbesondere durch die neue KI-Verordnung. Auch Forschungsorganisationen müssen zunehmend dokumentieren, welche Daten sie zu welchen Zwecken crawlen, insbesondere wenn die Grenze zur kommerziellen Verwertung überschritten wird. Die Pflicht zur Beachtung maschinenlesbarer Vorbehalte im Rahmen von § 44b UrhG wird durch die KI-Verordnung zu einem zentralen Baustein für den rechtssicheren Marktzugang von KI-Modellen in der EU.

Für die Hochschulpraxis empfiehlt es sich, klare Leitlinien für den Einsatz von Crawlern und das Training hochschuleigener Modelle zu etablieren. Da die Beweislast im Streitfall bei den Nutzenden liegt, ist eine lückenlose

technische Dokumentation der Erkennung von Opt-outs
unerlässlich. Letztlich wird die zukünftige Rechtsprechung klären müssen,
wie transformative Nutzungen im Rahmen des Crawlings gewichtet werden
und ob die derzeitigen Schrankenregelungen der Geschwindigkeit der
technischen Entwicklung standhalten können.

08 Literaturverzeichnis

- Ammann, Thorsten / Falk, Richard / Achnitz, Florian / Kurth, Sonja*, Der EU-Praxisleitfaden für GPAI-Modelle, Künstl. Intell. Recht 2026, [Link](#)
- Antoine, Lucien*, Die Pflichten aus Art. 53 I KI-VO – Hilfreich für das Urheberrecht?, Gewerbl. Rechtsschutz Urheberrecht 2025, [Link](#)
- Anton, Katharina*, Recht der elektronischen Medien, 2026,
- Baumann, Malte*, Generative KI und Urheberrecht – Urheber und Anwender im Spannungsfeld, Neue Juristische Wochenschr. 2023, 3673–3678
- Begerow, Markus*, Software Agents, 2025, [Link](#)
- Bomhard, David*, Urheberrecht, 2025,
- Creative commons*, About CC Licences, Creative commons [Link](#)
- Dreier, Thomas / Schulze, Gernot (Hrsg.)*, Urheberrechtsgesetz, 2025, [Link](#)
- Europäisches Parlament und Rat der Europäischen Union*, Verordnung (EU) 2022/2065 des Europäischen Parlaments und des Rates vom 19. Oktober 2022 über einen Binnenmarkt für digitale Dienste und zur Änderung der Richtlinie 2000/31/EG (Gesetz über digitale Dienste), ABl. L 2022, [Link](#)
- Gerecke, Martin*, Social Media und Recht: Einige urheberrechtliche Gedanken zu generativen KI-Modellen, Gewerbl. Rechtsschutz Urheberrecht Prax. 2023, [Link](#)
- Götting, Horst-Peter / Lauber-Rönsberg, Anne / Rauer, Nils*, Urheberrecht, 2025, [Link](#)
- Grimm, Susanne / Münster, Laura Marie*, Training von KI in der EU: Fragen zum Nutzungsvorbehalt beim Text und Data Mining sowie zur Beweislast, Gewerbl. Rechtsschutz Urheberrecht Prax. 2025, [Link](#)
- Jandt, Silke / Steidle, Roland (Hrsg.)*, Datenschutz im Internet, 2025,
- Käde, Lisa*, Know Your Data, Dtsch. Stift. Für Recht Inform. 2024, [Link](#)
- Käde, Lisa*, Training generativer KI-Modelle ist (auch) Text- und Data-Mining, Künstl. Intell. Recht 2024, 162–169

Kadek, Web Scraping vs. Web Crawling: Was ist der Unterschied?

Ein umfassender Vergleich, WebData Extraction [Link](#)

Knecht, Jana, Rechtsbereichsspezifische Betrachtung von KI: Urheberrecht, 2025, 1–27 [Link](#)

Leistner, Matthias, Urheberrecht und Künstliche Intelligenz, Gewerbl. Rechtsschutz Urheberrecht 2025, [Link](#)

LG München, LG München I: Unzulässige Vervielfältigung durch Memorisierung von Werken im und durch KI-Sprachmodell - GEMA/Open AI, Gewerbl. Rechtsschutz Urheberrecht 2025, [Link](#)

Maamar, Niklas, Urheberrechtliche Fragen beim Einsatz von generativen KI-Systemen, Z. Für Urh.- Medien. 2023, 481–492

Morgenroth, Carsten, Hochschulstudienrecht und Hochschulprüfungsrecht, 2026,

Nordemann, Axel / Nordemann, Jan Bernd / Czychowski, Christian (Hrsg.), Urheberrecht Fromm/Nordemann, 2024, [Link](#)

Pesch, Paulina Jo, Recht der elektronischen Medien, 2026,

Schippel, Robert, Creative-Commons-Lizenzen als Grenzen des KI-Trainings, Künstl. Intell. Recht 2026, 36–39

Schlotfeldt, Andrea, 10 Fragen zu KI-Output, Urheberrecht und OER, 2025, [Link](#)

Schmidt, Markus / Pruß, Michael, Handbuch IT- und Datenschutzrecht, 2019,

Wagner, Kristina, Generative KI: Eine „Blackbox“ urheberrechtlicher Haftungsrisiken?, Z. Für Recht Digit. Datenwirtschaft IT, 298–304

Wendt, Janine / Wendt, Domenik, Das neue Recht der Künstlichen Intelligenz, 2025,

Witte Paz, Annika / Seitz, Natalie, Künstliche Übersicht, echter Missbrauch? - Googles AI Overviews als Missbrauch von Suchmaschinen, NZKart 2025, [Link](#)

World Wide Web Consortium (W3C), TDM Reservation Protocol (TDMRep), 2024, [Link](#)

Ziegler, Florian / Jürges, Hannah, „Open-Source“ in der KI-Verordnung, RDi 2026 77 Ff 2026, 77–82

Kontakt

Rechtsinformationsstelle für die digitale Lehre
(bwDigiRecht)
im Hochschulnetzwerk Digitalisierung der Lehre
Baden-Württemberg (HND-BW)
Karlsruher Institut für Technologie (KIT)
Adenauerring 12
76131 Karlsruhe
bwDigiRecht@hnd-bw.de



Zitiervorschlag: Knecht, Jana, Handreichung Verhältnis von Data Crawling und Urheberrecht im Hochschulkontext, Rechtsinformationsstelle für die digitale Lehre (bwDigiRecht) im Hochschulnetzwerk Digitalisierung der Lehre Baden-Württemberg, Karlsruhe, 2026.

bwDigiRecht wird von den Hochschulen in Baden-Württemberg getragen



Diese Handreichung ist ausschließlich zur Nutzung durch die Trägerhochschulen bestimmt, eine Weitergabe an Dritte oder sonstige Verbreitung, ohne vorherige ausdrückliche Einwilligung der Koordination von bwDigiRecht, ist unzulässig.