Wie verhindere ich, dass meine Internetseiten zum KI-Training genutzt werden (dürfen)?

A. Überblick

1. Inhalte, die **unter CC0, CC-BY, CC-BY-SA, etc. lizenziert sind**, können zum Training benutzt werden (wobei die Attributionspflicht – d.h. die Nennung der Quelle – ggf. problematisch sein kann).[1] Dann spielt der Rechtevorbehalt (unten B) keine Rolle! Sie müssen also Inhalte, die Sie vor der Erfassung durch KI schützen wollen, unter eine andere Lizenz stellen (oder auf eine explizite Lizenz ganz verzichten).

2. Grundsätzlich dürfen **alle öffentlich zugänglichen Internetseiten** nicht nur durch Suchmaschinen, sondern auch zum Training von generativen KI-Systemen ausgewertet werden.[2] Wenn Ihre Inhalte also in einem geschlossenen System gespeichert sind (mit Anmeldung), können Sie den Nutzern auch konkrete Vorgaben für die Weiterverwendung machen und insb. die Auswertung zum KI-Training verbieten.

3. Eine Auswertung ist in jedem Fall ausgeschlossen, wenn ein **Rechtevorbehalt in maschinenlesbarer Form** vorliegt und die Inhalte überhaupt urheberrechtlich geschützt sind (also nicht ihrerseits maschinengeneriert, von Personen erstellt, die vor mehr als 70 Jahre gestorben sind oder von so geringer Schöpfungshöhe, dass sie nicht schutzfähig sind).

4. Gar nicht verhindern lässt sich die Auswertung öffentlicher Seiten zu wissenschaftlichen Zwecken – hier ist auch ein maschinenlesbarer Rechtevorbehalt nicht hinreichend (Art. 3 DSM-RL, § 60d UrhG).

---

[1] https://creativecommons.org/2023/08/18/understanding-cc-licenses-and-generative-ai/#:~:text=To%20develop%20and%20share%20principles,%20best%20practices,%20guidance.

[2] Zweifelnd an der Wirksamkeit der gesetzlichen Regelung https://www.lmr-nrw.de/aktuell/detail/initiative-urheberrecht-praesentiert-die-interdisziplinaere-studie-urheberrecht-training-generativer-ki-technologische-und-rechtliche-grundlagen#:~:text=Ihre%20interdisziplin%C3%A4re%20Forschung%20liefert%20dringend%20ben%C3%B6tigte.

B. Wie formuliere ich einen „Rechtevorbehalt in maschinenlesbarer Form“?

Variante 1: robots.txt

Man kann sog. Crawler davon abhalten, Webseiten automatisiert zu erfassen. Dazu gibt es das sog. **Robots Exclusion Protocol**.[3] Man muss dazu eine Datei mit dem Namen „robots.txt“ im Hauptverzeichnis einer Internetdomain ablegen. Die Datei liegt dann z.B. unter https://jura.uni-passau.de/robots.txt.

In dieser Datei kann man dann für einzelne Ordner oder Dateien die Auswertung durch alle oder bestimmte Crawler verbieten bzw. erlauben.

Enthält die Datei die beiden Zeilen

```
User-agent: *
Disallow: /
```

wird allen automatisierten Programmen die Auswertung **aller Inhalte der Webseite** verboten. Dies umfasst aber z.B. auch Suchmaschinen wie Google und Bing, welche die Inhalte dann nicht mehr erfassen.

Schlauer ist es also, **bestimmte Crawler** auszuschließen. OpenAI (ChatGPT)[4] und Google[5] haben entsprechende Listen bereitgestellt. Eine weitergehende und laufend aktualisierte Auflistung gibt es unter **https://raw.githubusercontent.com/ai-robots-txt/ai.robots.txt/refs/heads/main/robots.txt**

Wenn man diese Datei herunterlädt und **im Hauptverzeichnis des eigenen Servers speichert**, darf die komplette Website nicht von KI-Systemen erfasst werden.

Sollen **nur bestimmte Verzeichnisse ausgeschlossen** werden, muss die letzte Zeile „Disallow: /“ ersetzt werden durch eine Liste der Verzeichnisse, die ausgeschlossen sein sollen. Also zum Beispiel

```
Disallow: /geheim/
Disallow: /nichtfuerki/
```

etc.

Für **Wordpress** gibt es ein Plugin (https://wordpress.org/plugins/wp-robots-txt/), mit dem Sie diese Datei bearbeiten können.

---

3 https://datatracker.ietf.org/doc/html/rfc9309.

4 https://platform.openai.com/docs/bots.

5 https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers.

Variante 2: TDMRep

Ebenfalls möglich ist die Verwendung des sog. TDMRep-Protokolls.[6] Dazu muss im HTML-Code im Header der Internetseite folgende Zeile ergänzt werden, um den Zugriff zu verbieten:

```
<meta name="tdm-reservation" content="1">
```

Will man den Zugriff unter bestimmten Bedingungen erlauben, kann man das auf einer Internetseite ausformulieren oder als JSON-Datei kodieren, letzteres wird in der Protokollspezifikation genau ausformuliert.

Für **Wordpress** gibt es hierzu bereits ein Plugin (https://wordpress.org/plugins/tdmrep/), welches dies benutzerfreundlich ermöglicht.

Der Nachteil von TDMRep ist, dass es derzeit nur ein Vorschlag ist und noch nicht überall automatisiert erkannt wird.

C. Rechtliche Grundlagen

Art. 4 DSM-RL 2016/790- Ausnahmen und Beschränkungen für das Text und Data Mining

(1) Für zum Zwecke des Text und Data Mining vorgenommene Vervielfältigungen und Entnahmen von rechtmäßig zugänglichen Werken und sonstigen Schutzgegenständen sehen die Mitgliedstaaten eine Ausnahme oder Beschränkung von den Rechten vor, die in Artikel 5 Buchstabe a und Artikel 7 Absatz 1 der Richtlinie 96/9/EG, Artikel 2 der Richtlinie 2001/29/EG, Artikel 4 Absatz 1 Buchstaben a und b der Richtlinie 2009/24/EG und Artikel 15 Absatz 1 der vorliegenden Richtlinie niedergelegt sind.

(2) Vervielfältigungen und Entnahmen nach Absatz 1 dürfen so lange aufbewahrt werden, wie es für die Zwecke des Text und Data Mining notwendig ist.

(3) Die Ausnahmen und Beschränkungen nach Absatz 1 finden Anwendung, **sofern die jeweiligen Rechteinhaber die in Absatz 1 genannten Werke und sonstigen Schutzgegenstände nicht ausdrücklich in angemessener Weise, etwa mit maschinenlesbaren Mitteln im Fall von online veröffentlichten Inhalten, mit einem Nutzungsvorbehalt versehen haben**. …

[6] https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240202/.

Erwägungsgrund 18 DSM-RL 2016/790[7]

... Wurden Inhalte **im Internet öffentlich zugänglich gemacht**, so sollte es als angemessen erachtet werden, einen **Rechtsvorbehalt mit maschinenlesbaren Mitteln** auszusprechen; Das gilt **auch für Metadaten und Geschäftsbedingungen einer Website oder eines Dienstes**. Andere Nutzungen sollten von dem Rechtsvorbehalt für die Zwecke des Text und Data Mining nicht betroffen sein. In anderen Fällen kann es angemessen sein, einen Rechtsvorbehalt mit anderen Mitteln, etwa in vertraglichen Vereinbarungen oder durch eine einseitige Erklärung, auszusprechen. Die Rechteinhaber sollten in der Lage sein, Maßnahmen zu treffen, mit denen sie sicherstellen, dass ihre diesbezüglichen Vorbehalte Beachtung finden. ...

§ 44b UrhG - Text und Data Mining

(1) Text und Data Mining ist die automatisierte Analyse von einzelnen oder mehreren digitalen oder digitalisierten Werken, um daraus Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen.

(2) Zulässig sind Vervielfältigungen von rechtmäßig zugänglichen Werken für das Text und Data Mining. Die Vervielfältigungen sind zu löschen, wenn sie für das Text und Data Mining nicht mehr erforderlich sind.

(3) **Nutzungen nach Absatz 2 Satz 1 sind nur zulässig, wenn der Rechtsinhaber sich diese nicht vorbehalten hat. Ein Nutzungsvorbehalt bei online zugänglichen Werken ist nur dann wirksam, wenn er in maschinenlesbarer Form erfolgt.**

[7] 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 (Text von Bedeutung für den EWR).