Von CC bis OSAID: Freie Lizenzen für Daten, Datenbanken und Datenmodelle

Freiheit für die Daten

Freie Lizenzen werden nicht nur für Software genutzt, sondern auch in anderen Bereichen. Für Daten, Datenbanken und LLMs gibt es spezialisierte Varianten. Frank Hofmann, Veit Schiele

Die Autoren

Frank Hofmann arbeitet zumeist unterwegs als Entwickler, Trainer und Autor, bevorzugt in Berlin, Genf und Kapstadt. Er gehört zu den Verfassern des Debian-Paketmanagement-Buchs. **Veit Schiele** ist Gründer und Geschäftsführer des Technologieberatungsunternehmens Cusy GmbH. Es verwendet Prozesse, die das Einhalten von Rechtsvorschriften automatisiert überprüfen. Veit ist zudem Autor des Tutorials „Python for Data Science".

Dieser Artikel ist der letzte Teil unserer fünfteiligen Serie über freie Lizenzen. In den ersten drei Beiträgen bezogen wir uns auf Software. Der vierte Teil legte den Schwerpunkt auf Dokumente, Bilder, Audio- und Videodaten sowie Schriftarten und Hardware. Im Folgenden sehen wir uns an, wie Lizenzangaben bei Daten, Datenbanken und Datenmodellen aussehen können.

Überblick

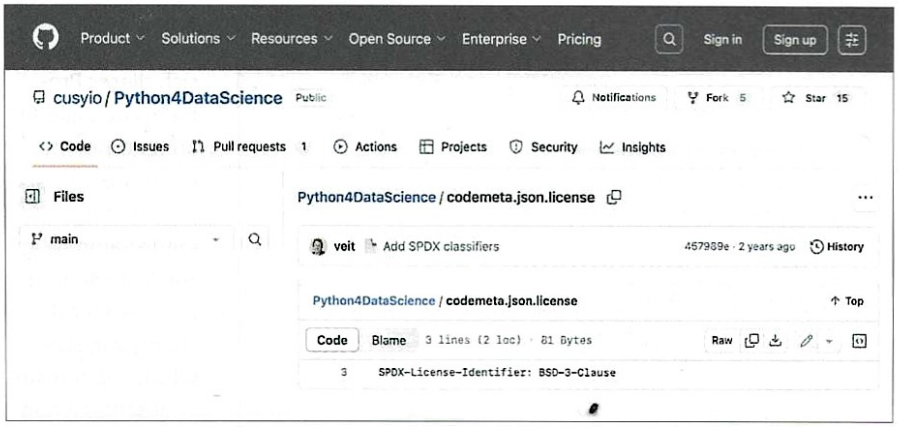
Daten, Datenbanken und Datenmodelle lassen sich nicht mit den üblichen Open-Source-Lizenzen frei verfügbar machen. Es gibt für diesen Zweck jedoch spezielle Lizenzen, die eine freie Nutzung ermöglichen. Damit wuchs auch der Druck auf öffentliche Verwaltungen und Regierungen, die Datengrundlage ihrer Entscheidungen offenzulegen.

Die Tabelle Lizenzen für Daten, Datenbanken und Datenmodelle vermittelt einen Überblick zu freien Nicht-Software-Lizenzen. Nennen wir darin zu einer Lizenz lediglich Creative Commons (CC-BY-*-4.0), heißt das, dass es mehrere Varianten gibt und wir als Lizenzgeber eine genauere Auswahl innerhalb der Kategorie treffen müssen. Die Tabelle liefert jedoch nur eine erste Übersicht. Im Folgenden gehen wir noch genauer auf die einzelnen Lizenzen ein und zeigen, wie Sie die Angaben in unterschiedlichen Dateiformaten machen können.

Lizenzangaben machen

Es gibt mehrere Möglichkeiten, um Autorenschaft, Lizenz und Link zum Lizenzdokument mitzuteilen: etwa in der Nähe der Datei, in den Metainformationen der Datei, in den Daten selbst oder in einer separaten Datei. Listing 1 zeigt ein Beispiel für die Lizenzangabe in der Datei selbst, im konkreten Fall im JSON-Format in Kombination mit SPDX für zwei Grafiken im SVG-Format. Beide Grafiken sind unter Creative Commons mit Namensnennung des Autors lizenziert.

Abbildung 1 demonstriert das Vorgehen in einer separaten Datei. SPDX-konform ist eine zusätzliche, passende `*.license`-Datei zum Projekt oder Werk, im Beispiel für eine JSON-Datei aus dem Python4DataScience-Tutorial. Dank der Maschinenlesbarkeit der SPDX-Angabe lassen sich mögliche Lizenzkonflikte automatisiert erkennen.



1 Eine SPDX-Lizenzangabe für eine JSON-Datei aus dem Python4DataScience-Tutorial.

Daten(banken) und LLMs

Ähnlich wie bei Dokumenten lassen sich bei Daten, Datenbanken und Datenmodellen die Lizenzen in den Daten selbst angeben oder als Verweis mitliefern, beispielsweise via SPDX.

Autoren und Lizenzen speichern Sie einfach zusammen mit den Daten in der Datenbank. Dabei sollten Sie eine Datenlizenz wählen, die sich von der Lizenz des Datenbankmanagementsystems (DBMS) unterscheidet. Die Nutzer können die Lizenzinformationen über die Abfragesprache des DBMS maschinell auslesen.

Die BigScience Open RAIL-M License kann zusammen mit Datenmodellen wie LLMs ausgeliefert werden. Das Debian-Projekt und die Open Source Initiative (OSI) erkennen KI-Modelle allerdings nur als frei an, wenn sie zusammen mit den Trainingsdaten und -programmen bereitgestellt werden.

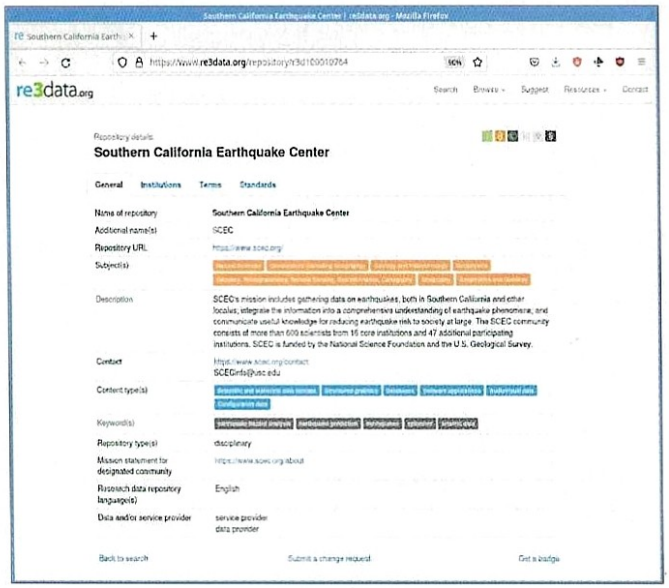
Listing 1: Referenzliste im JSON-Format

```
{
  "licenses": {
    "balkendiagramm.svg": "CC BY-SA 4.0",
    "tortendiagramm.svg": "CC BY-SA 4.0"
  }
}
```

Lizenzen für Daten, Datenbanken und Datenmodelle

Lizenz	SPDX-Angabe
Daten	
Creative Commons (CC0)	CC0-1.0
Datenlizenz Deutschland – Zero – Version 2.0	DL-DE-ZERO-2.0
Datenlizenz Deutschland – Namensnennung – Version 2.0	DL-DE-BY-2.0
Open Data Commons Attribution License (ODC-By v1.0)	ODC-By-1.0
Open Data Commons Public Domain Dedication and License (PDDL v1.0)	PDDL-1.0
Community Data License Agreement Permissive 2.0 (CDLA2 [58])	CDLA-Permissive-2.0
Community Data License Agreement Sharing 1.0 [59]	CDLA-Sharing-1.0
Computational Use of Data Agreement v1.0	C-UDA-1.0
Open Use of Data Agreement v1.0	O-UDA-1.0
Datenbanken	
Open Data Commons Open Database License (ODbL v1.0)	ODbL-1.0
Creative Commons (CC0)	CC0-1.0
Creative Commons International 4.0	CC-BY-*-4.0
Datenmodelle	
BigScience Open RAIL-M License	–
OSAID 1.0	–

Werden Sie KI-Profi!
Der KI-Newsletter des Linux-Magazins
www.linux-magazin.de/subscribe



2 Offene Forschungsdaten, wie hier die des Southern California Earth Quake Center, finden Sie bei Re3data.

Welche Lizenz?

Zur Beantwortung der Frage, welche Lizenz sich denn nun für die eigenen Daten am besten eignet, gibt es für die meisten Datentypen eine Reihe von Hilfestellungen. Dazu zählen unter anderem der License Chooser (Creative Commons), der Guide to Open Data Licensing sowie die Website Choose A License. Die rechtlichen Hintergründe erörtern das iRights-Institut sowie ifrOSS, das Institut für Rechtsfragen der freien und Open-Source-Software. Datenmodelle behandeln diese beiden Quellen jedoch nicht.

Offene Daten dürfen von jedermann zu jedem Zweck genutzt und verbreitet werden. Nutzungseinschränkungen sind nur erlaubt, um Ursprung und Offenheit des Wissens zu sichern, sofern es sich um geistiges Eigentum handelt. Somit veröffentlichen viele Institute und Forschungseinrichtungen sowohl ihre Forschungsdaten als auch die dazugehörigen Modelle zur Berechnung. Entsprechende Daten finden Sie bei Interesse beispielsweise über Re3data, die Registry of Research Data Repositories 2.

Sofern Daten lediglich eine Zahlenmenge darstellen, gibt es ohnehin keine Urheberschaft. Erst durch die Erfassung und Interpretation erhalten sie ihren urheberrechtlichen Schutz. Das Research Data Repository (RADAR) hat sich dieses Problems detaillierter angenommen. RADAR, eine disziplinenübergreifende Sammlung zur Archivierung und Veröffentlichung von Forschungsdaten aus abgeschlossenen wissenschaftlichen Studien und Projekten, hat ihren Sitz am FIZ Leibniz-Institut für Informationsinfrastruktur in Karlsruhe.

Hinsichtlich der Wahl einer Lizenz für Forschungsdaten rät RADAR zu einer der Varianten der Creative-Commons-Lizenzen. Zudem empfiehlt das Institut drei Schritte zur Überprüfung der Daten vor der Veröffentlichung:

- Die Inhaber der Rechte an den Forschungsdaten lassen sich eindeutig identifizieren, und es liegt die Berechtigung vor, die Daten öffentlich oder für einen bestimmten Nutzerkreis zugänglich zu machen.
- Der potenzielle kommerzielle Wert der Daten wurde bedacht. Bereits veröffentlichte Forschungsdaten können zum Beispiel in der Regel nicht mehr im Rahmen einer Patentanmeldung verwendet werden.
- Eine möglichst weitreichende Nachnutzung der Forschungsdaten liegt im Sinne der Datengeber beziehungsweise -autoren.

Das Konzept hinter Open Data geht auf Überlegungen anlässlich des Internationalen Geophysikalischen Jahrs 1957/58 zurück. Es ging darum, die Metadaten zu standardisieren, um den Austausch und die Nutzung wissenschaftlicher Daten zu erleichtern. Aus dieser Motivation entstand die Open-Access-Bewegung, die Forschungsergebnisse der Allgemeinheit frei zugänglich machen möchte.

Die Idee von Open Data geht mittlerweile weit über wissenschaftliche Daten hinaus und findet sich etwa in Konzepten wie Open Government und den Informationsfreiheitsgesetzen wieder. Im Zusammenhang mit Open Government entstand das Datenportal für Deutschland (GovData) mit der Datenlizenz Deutschland, einmal ohne und einmal mit Namensnennung.

Weitere Lizenzen für Daten sind die Data-Commons-Lizenzen mit der Open Data Commons Attribution License (ODC-By v1.0) und der Open Data Commons Public Domain Dedication and License (PDDL v1.0).

Datenbanklizenzen

Man kann nicht nur einzelne Daten unter einer freien Lizenz veröffentlichen, sondern auch komplette Datenbanken samt

3 Ein Kartenausschnitt aus OpenStreetMap. Die Daten dazu unterliegen der ODbL v1.0.

der zugehörigen Strukturbeschreibungen. Der Gesetzgeber unterscheidet hier zwischen Datenbankwerken und Datenbankherstellerrecht (sogenannte Sui-generis-Datenbanken).

Ein Datenbankwerk kann urheberrechtlich geschützt werden, wenn „aufgrund der Auswahl oder Anordnung der Elemente eine persönliche geistige Schöpfung" vorliegt (§ 4 UrhG). Das Datenbankherstellerrecht entsteht hingegen, wenn die Beschaffung oder Darstellung auf einer wesentlichen Investition beruht (§ 87a UrhG). Als Lizenzen eignen sich hier CC0, die Version 4 der Creative-Commons-Lizenzen sowie die Open Data Commons Open Database License (ODbL v1.0). Letztere verwendet beispielsweise OpenStreetMap 3.

Spezialfall KI

Umstritten ist, ob die Unterscheidung bei der Lizenzierung von Datenbanken auch für Modelle des maschinellen Lernens gilt und ob sich diese überhaupt lizenzieren lassen. Sofern man sie als mathematische Formeln ausdrückt, ähneln sie einem Naturprinzip. Handelt es sich hingegen um wissenschaftliche oder technische Methoden beziehungsweise Entdeckungen, kann man sie beispielsweise unter die BigScience Open RAIL-M License stellen. RAIL steht dabei für Responsible AI License und zielt auf verhaltensbasierte Nutzungsbeschränkungen ab, um das Risiko von Schäden durch die gemeinsame Nutzung zu verringern.

Vorsicht ist bei der Nutzung von Metas Sprachmodell Llama geboten: Meta deklariert es als Open Source, obwohl die verwendeten Lizenzvereinbarungen weder von der OSI anerkannt wurden noch einer freien Lizenz entsprechen. Vermutlich möchte Meta damit einer schärferen Kontrolle durch den EU AI Act entgehen, der spezielle Regeln für Open-Source-Modelle enthält, ohne die Einhaltung der OSI-Richtlinien zu verlangen.

Das Debian-Projekt hat sich ebenfalls in Bezug auf künstliche Intelligenz positioniert: Seit Mai 2025 gelten KI-Modelle, die unter einer Open-Source-Lizenz stehen und ohne Schulungsdaten und Programm veröffentlicht wurden, als nicht konform mit den Debian Free Software Guidelines (DFSG).

© Armel Le Coz, Cyril Lage, CC BY 3.0: https://commons.wikimedia.org/w/index.php?curid=57401884

4 Das Kreislaufmodell von DémocratieOuverte.org für eine offene Demokratie.

Auch für die OSI geht die Definition von Open-Source-KI weit über das bloße Verwenden eines Modells hinaus: Die zum Training des Systems verwendeten Daten müssen verfügbar sein, um ein gleichwertiges System aufbauen zu können. Der Quellcode für das Training und den Betrieb des Systems muss unter von der OSI genehmigten Lizenzen veröffentlicht sein. Die Modellparameter, Gewichte oder andere Konfigurationseinstellungen müssen ebenfalls unter von der OSI genehmigten Lizenzen vorliegen.

Die OSI entwickelte entsprechend die Lizenz OSAID 1.0, die unter anderem für folgende Modelle gilt:

- EleutherAI: Pythia und GPT-J
- The Allen Institute for Artificial Intelligence: OLMo 2 und Molmo
- LLM360: K2, Amber und CrystalCoder
- Google: T5

Vermutlich würden auch die folgenden Modelle die Anforderungen erfüllen, sofern sie ihre rechtlichen Bedingungen ändern würden:

- BigScience: Bloom
- BigCode: StarCoder 2
- Technology Innovation Institute: Falcon

Es gibt jedoch auch einige Modelle, die von der OSI analysiert wurden und den

Ideen der Open-Government-Bewegung

Titel	bisher	neu
Öffentlichkeit und Geheimhaltung	Alles, was nicht ausdrücklich als öffentlich gekennzeichnet ist, bleibt geheim.	Alles, was nicht ausdrücklich als geheim gekennzeichnet wurde, ist öffentlich.
Umfang, Art und Zeitpunkt der Veröffentlichung	Die einzelnen Behörden bestimmen darüber selbst.	Die Behörden veröffentlichen proaktiv, zeitnah und vollumfänglich, sofern die Daten keinem berechtigten Datenschutz oder Datensicherheitsbeschränkungen unterliegen.
Nutzungsrechte	Veröffentlichte Daten werden nur für den privaten Gebrauch zur Einsicht freigegeben, weitere Nutzungsrechte werden nur fallweise gewährt.	Jedermann darf die veröffentlichten Daten für jegliche Zwecke kostenfrei nutzen.

```
>>> from deutschland.lebensmittelwarnung import Lebensmittelwarnung
>>> lw = Lebensmittelwarnung()
>>> data = lw.get("lebensmittel", "berlin")
>>> print(data[0])
{'guid': 'https://www.lebensmittelwarnung.de/___lebensmittelwarnung.de/Meldungen
/2025/08_August/250801_02_NW_Salami/250801_02_NW_Salami.html', 'pubDate': 'Fri,
1 Aug 2025 12:10:00 +0200', 'description': '<img src="https://www.lebensmittelwa
rnung.de/___lebensmittelwarnung.de/Meldungen/2025/08_August/250801_02_NW_Salami/
250801_02_NW_Salami_Bild.png?__blob=normal&amp;v=2" width="100" /><br/><b>Bildqu
elle</b> © Kundeninformation der Firma Franz Wiltmann GmbH & Co. KG<br/><b>Verpa
ckungseinheit:</b> 70 Gramm<br/><b>Chargennummer / Los-Kennzeichnung:</b> L25177
00010<br/><b>Grund der Meldung:</b>  Krankheitserreger<br/><b>Haltbarkeit:</b>
14.08.2025, 21.08.2025, 25.08.2025<br/><b>Hersteller / Inverkehrbringer:</b> Fra
nz Wiltmann GmbH & Co. KG\nW.-Kleine-Straße 16\n33775 Versmold<br/><b>Betroffene
 Bundesländer nach derzeitigem Stand:</b>  Baden-Württemberg, Bayern, Berlin, B
randenburg, Bremen, Hessen, Mecklenburg-Vorpommern, Niedersachsen, Nordrhein-Wes
tfalen, Rheinland-Pfalz, Saarland, Sachsen, Sachsen-Anhalt, Schleswig-Holstein,
Thüringen<br/>', 'link': 'https://www.lebensmittelwarnung.de/___lebensmittelwarn
ung.de/Meldungen/2025/08_August/250801_02_NW_Salami/250801_02_NW_Salami.html', '
title': 'Rein Rind Salami, 70 Gramm'}
```

5 **Eine über die Python-Bibliothek Deutschland ausgelesene Lebensmittelwarnung.**

Kriterien für eine freie Lizenz nicht genügen, weil ihnen erforderliche Komponenten beziehungsweise rechtliche Vereinbarungen fehlen:

- Meta: Llama2
- xAI: Grok
- Microsoft: Phi-2
- Mistral AI: Mixtral

Am 5. August 2025 veröffentlichte OpenAI die Open-Weight-Modelle Gpt-oss-120b und Gpt-oss-20b, die sich für lokale Inferenzen und schnelle Iterationen ohne teure Infrastruktur eignen. Der Hersteller gab sogar an, wie die Modelle trainiert wurden, und stellte darüber hinaus OpenAI Harmony unter der Apache-2.0-Lizenz bereit. Open Weights führen zwar zu mehr Transparenz in den KI-Modellen und ermöglichen damit einen besseren Einblick in die Funktionsweise neuronaler Netze. Das genügt aber nicht, um solche Modelle als OSI Approved License zu klassifizieren.

Stellt der Lizenzgeber weder den Trainingscode noch die Trainingsdatensätze offen zur Verfügung, lässt sich nur der Endzustand des Modells analysieren, nicht jedoch der Trainingsprozess. Damit ist eine der Voraussetzungen wissenschaftlichen Arbeitens nicht gegeben, potenzielle Verzerrungen in den Trainingsdaten bleiben verborgen. In Trainingsdaten, die nicht repräsentativ und ethisch einwandfrei sind, schleicht sich allzu leicht Garbage In, Garbage Out ein.

Regierungen weltweit formulieren Richtlinien, die transparentere KI-Modelle vorschreiben. Mit Open Weights versuchen die KI-Anbieter möglicherweise, diese Richtlinien zu erfüllen, ohne den Schulungscode und die Trainingsdaten offenlegen zu müssen.

Open Government

Die Idee, Software frei zur Verfügung zu stellen, wurde in den letzten Jahrzehnten auf viele weitere Bereiche übertragen, darunter Daten, Datenbanken und Datenmodelle. Das wirkt sich auch auf öffentliche Verwaltungen und Regierungen aus. Seit 2011 versucht der Arbeitskreis Open Government Netzwerk Deutschland die Bundesregierung dazu zu bewegen, der internationalen Initiative Open Government Partnership (OGP) beizutreten, die 2011 von den Regierungen der USA und Brasiliens ins Leben gerufen wurde. 2013 vereinbarten Union und SPD im Koalitionsvertrag die Teilnahme an der OGP, 2016 leitete die Regierung sie mit der Überreichung der Absichtserklärung offiziell ein.

Die Open-Government-Bewegung möchte einen Paradigmenwechsel herbeiführen, die Regierung und Verwaltung öffnen und auf diese Weise mehr Transparenz, Teilhabe, Zusammenarbeit und Innovation erzielen. Daraus ergibt sich ein Kreislauf, der allen nutzt **4**. Die Tabelle Ideen der Open-Government-Bewegung (vorige Seite) stellt die angestrebten Veränderungen gegenüber. 2013 ging das Datenportal für Deutschland (GovData) online, das auf offene Verwaltungsdaten von Bund, Ländern und Kommunen verweist. Sie unterliegen den bereits oben genannten Lizenzen für Daten, was für eine deutliche Öffnung und einen vereinfachten Umgang mit den Verwaltungsdaten geführt hat.

Im Juni 2021 wurde das zweite Open-Data-Gesetz beschlossen. Unverständlich bleibt jedoch, dass offiziell weiter keine offenen Programmierschnittstellen (APIs) bereitgestellt werden müssen. Daher betrachten viele das Gesetz als Papiertiger und befürchten, dass die Daten der öffentlichen Verwaltungen und Regierungen weiterhin schwer zugänglich bleiben. So hat Lilith Wittmann zwei Monate nach Veröffentlichung des Gesetzes eine Webseite erstellt, die die Dokumentation offener Programmierschnittstellen (APIs) von öffentlichen Behörden zugänglich macht. Ergänzend gibt es ein Git-Repository für Software, die diese APIs nutzt **5**.

Seit Juni 2023 gibt es die Plattform FragDenStaat, die wichtige rechtliche und parlamentarische Dokumente aufbereitet und frei zur Verfügung stellt. Dazu zählen unter anderem das Gemeinsame Ministerialblatt, das Bundessteuerblatt und das Verkehrsblatt. Daneben finden sich hier kleine Anfragen aus deutschen Parlamenten sowie aus dem Bundestag. Hinzu kommen Gutachten der wissenschaftlichen Dienste sowie Dokumente aus den Untersuchungsausschüssen und dem Vermittlungsausschuss.

Fazit

Die Autoren freuen sich, dass die Ideen hinter quelloffener und freier Software zunehmend in anderen gesellschaftlichen Bereichen Einzug halten und damit auch öffentliche Verwaltungen und Regierungen zunehmend transparenter werden. Das hilft dabei, die Zivilgesellschaft weiter zu stärken. (*jlu*)

Danksagung

Die Autoren bedanken sich bei Werner Heuser für dessen Kritik und Anregungen bei der Vorbereitung des Artikels.

Weitere Infos und interessante Links

www.lm-online.de/qr/52528