Was ist Common Crawl? Eine Geschichte des offenen Web-Datensatzes

By rankstudio.net Published October 27, 2025 48 min read

Zusammenfassung

Common Crawl ist eine **501(c)(3)** gemeinnützige Stiftung (gegründet 2007), die ein **kostenloses, offenes Repository für Web-Crawl-Daten** pflegt (Source: commoncrawl.org) (Source: commoncrawl.org). Ihre Mission ist es, *den Zugang zu Webinformationen zu demokratisieren*, indem sie **Web-Crawl-Datensätze im Petabyte-Bereich** kostenlos zur Verfügung stellt. In den letzten über 15 Jahren hat Common Crawl in der Größenordnung von **300–400 Milliarden Webseiten** gesammelt, die mehr als 15 Jahre kontinuierlichen Crawlings umfassen (Source: commoncrawl.org) (Source: www.96layers.ai). Jeden Monat kommen etwa **3–5 Milliarden neue Seiten** hinzu (ca. 90 TB komprimiert, ~400 TB unkomprimiert) (Source: www.96layers.ai) (Source: commoncrawl.org). Obwohl es als winziges Projekt begann (nur wenige Mitarbeiter) (Source: www.96layers.ai), bildet Common Crawls öffentlich zugänglicher Korpus heute die Grundlage für eine Vielzahl von Forschungs- und kommerziellen Anwendungen. Insbesondere liefert er den Großteil der Trainingsdaten für moderne große Sprachmodelle (LLMs) – zum Beispiel stammten **über 80 % der Tokens in OpenAI's GPT-3** aus Common Crawl-Daten (Source: www.mozillafoundation.org) – und wird in über **10.000 wissenschaftlichen Publikationen** zitiert (Source: commoncrawl.org) (Source: dallascard.github.io). Es hat Startups (z.B. TinEye, Lucky Oyster) und Forschungsprojekte (z.B. GloVe Word Embeddings, Web-Zensuranalyse) ermöglicht, denen sonst die Ressourcen gefehlt hätten, das gesamte Web zu crawlen. Common Crawl dient somit als „neutrale, gemeinnützige Infrastruktur" für Webdaten (Source: www.96layers.ai), wodurch das Spielfeld geebnet wird, sodass selbst kleine Organisationen und Forscher auf webweite Informationen zugreifen können.

Dieser Bericht bietet eine **umfassende Geschichte und Analyse von Common Crawl**. Er behandelt die Ursprünge des Projekts (Schlüsselmotivationen, Gründerhintergrund, frühe Entwicklung), die Organisationsstruktur und Finanzierung, Datenerfassungsmethoden und -technologie, das Wachstum des Datensatzes und die **vielfältigen Einsatzmöglichkeiten der Daten heute** (im KI-/LLM-Training, in der akademischen Forschung, in Industrieprodukten usw.). Wir werden den sozialen und technischen Kontext untersuchen (z.B. die Dominanz von Google und die Notwendigkeit offener Webdaten), **quantitative**

Statistiken zusammenfassen (gesammelte Seiten, Datenvolumen, Zitationszahlen) und Fallstudien präsentieren, die den Einfluss von Common Crawl veranschaulichen. Wir diskutieren auch Herausforderungen (Abdeckungsbias, Urheberrechtsprobleme) und zukünftige Richtungen. Alle Behauptungen und Fakten werden durch maßgebliche Quellen der Common Crawl-Organisation, Medien, Interviews und Forschungspublikationen gestützt.

Einführung und Hintergrund

Das **World Wide Web** hat sich zu einem riesigen, dezentralen Informationsökosystem entwickelt. Moderne Suchmaschinen wie Google und Bing crawlen das Web kontinuierlich, um ihre eigenen Indizes zu erstellen, aber diese Indizes sind proprietär. Mitte der 2000er Jahre existierte **kein großes, öffentlich zugängliches Repository für Web-Crawl-Daten** für Außenstehende. Nur wenige Organisationen – insbesondere das gemeinnützige Internet Archive – versuchten, Webseiten zu archivieren (z.B. über die Wayback Machine). Die *Wayback Machine* des Internet Archive ist jedoch für die On-Demand-Snapshot-Archivierung und das Browsen von Webseiten über die Zeit konzipiert; sie ist nicht für groß angelegte Datenanalyse oder algorithmisches Mining des Webinhalts optimiert (Source: dallascard.github.io).

In diesem Kontext begann die Idee, einen **„offenen Web-Index“** aufzubauen, aufzukommen. Unternehmer und Forscher erkannten, dass nur die größten Unternehmen (Google, Microsoft, Yahoo, Baidu usw.) die Ressourcen hatten, Milliarden von Seiten mit hoher Frequenz zu crawlen, wodurch kleinere Akteure keinen Zugang zu diesen Rohdaten hatten. Zum Beispiel benötigten Universitätsforscher und Startups oft große Webkorpora für Aufgaben der natürlichen Sprachverarbeitung (NLP), des Data Mining und des maschinellen Lernens, aber es fehlten ihnen die Mittel, das gesamte Web selbst zu crawlen. Ein offenes Repository für Web-Crawl-Daten würde den **Zugang demokratisieren** und Innovationen fördern, ähnlich wie offene Datensätze (z.B. Wikipedia) neue Forschung vorantrieben.

Common Crawl wurde konzipiert und ins Leben gerufen, um diesem Bedarf gerecht zu werden. Sein Gründer, **Gil Elbaz**, ist ein Serienunternehmer und Technologe: Ende der 1990er Jahre war er Mitbegründer von Applied Semantics (dem Unternehmen, das die später als Google AdSense bekannte Technologie entwickelte) (Source: www.96layers.ai) (Source: www.96layers.ai). Nachdem Google Applied Semantics übernommen hatte, arbeitete Elbaz bis 2007 bei Google. In Interviews erklärte er, dass sein Weggang durch die Sorge über die Datenkonzentration und deren Auswirkungen auf die Innovation motiviert war. Er sah Googles massiven proprietären Crawl als Schlüssel zu seinem Monopol auf Suchinnovation (Source: www.96layers.ai) (Source: www.96layers.ai). Um dem entgegenzuwirken, stellte sich Elbaz „neutrale Datenunternehmen“ vor – offene, gemeinnützige Infrastrukturprojekte, die das Web crawlen und die Daten **kostenlos** Forschern und Unternehmen zur Verfügung stellen würden. Eines dieser Projekte war **Common Crawl**, gegründet 2007. Wie Elbaz es ausdrückte:

> *„Common Crawl sollte wie eine neutrale, gemeinnützige Infrastruktur sein, die die Art und Weise imitieren sollte, wie Google das Web crawlt … und diese Daten dann jedem kostenlos zur Verfügung stellen, um das Spielfeld der Technologieentwicklung zu ebnen“ (Source: www.96layers.ai).*

Elbaz’ Motivation war es daher explizit, **das Spielfeld zu ebnen**. Er wollte, dass kleine Startups und akademische Forscher die gleichen rohen „Suchindex“-Informationen wie Google hatten – damit Innovation nicht von einem Unternehmen monopolisiert würde (Source: www.novaspivack.com) (Source: www.96layers.ai). Diese Vision fand Anklang bei anderen Führungspersönlichkeiten der Open-Web-Community. Prominente Technologen wie **Nova Spivack** (ein früher Internet-Unternehmer) und **Carl Malamud** (ein Pionier offener Regierungsdaten) traten dem Gründungsbeirat von Common Crawl bei (Source: www.novaspivack.com). Im Laufe der Zeit wuchs der Beirat um Koryphäen wie Google-Forschungsdirektor **Peter Norvig** und MIT Media Lab-Direktor **Joi Ito** (Source: www.thekurzweillibrary.com) (Source: commoncrawl.org), was die Bedeutung des Projekts unterstreicht.

Innerhalb weniger Jahre wurde Common Crawl eine unabhängige gemeinnützige Stiftung. Bei ihrer Gründung wurde sie als kalifornische 501(c)(3)-Organisation, die **Common Crawl Foundation**, registriert (Source: commoncrawl.org) (Source: commoncrawl.org). Ihr Leitbild ist prägnant: *„den Zugang zu Webinformationen durch die Erstellung und Pflege eines offenen Web-Crawls zu demokratisieren“*. Die Common Crawl-Homepage beschreibt es als *„ein kostenloses, offenes Repository für Web-Crawl-Daten, das von jedem genutzt werden kann.“* (Source: commoncrawl.org). Gil Elbaz fungierte als Vorsitzender des Vorstands und wird oft als Gründer des Projekts genannt (Source: commoncrawl.org) (Source: www.novaspivack.com). Zu den weiteren wichtigen frühen Teammitgliedern gehörten der leitende Ingenieur **Ahad Rana** und später die Direktorin **Lisa Green** (ehemals Creative Commons) (Source: www.novaspivack.com).

Organisationsstruktur und Finanzierung

Common Crawl agiert als kleine gemeinnützige Organisation. Ihre Homepage und Teamseiten von 2025 weisen darauf hin, dass das Kernteam historisch sehr klein war – buchstäblich „weniger als fünf Personen“ in den frühen Jahren (Source: www.96layers.ai). Zum Beispiel wurde das Projekt in den frühen 2010er Jahren mit nur einer Handvoll Ingenieuren und Freiwilligen betrieben. Selbst als OpenAI 2020 das GPT-3-Paper veröffentlichte, hatte Common Crawl Berichten zufolge nur **einen Vollzeitmitarbeiter** (Source: www.96layers.ai) (obwohl das Team bis 2025 größer ist). Gil Elbaz fungiert als Vorsitzender (und war Co-Vorsitzender von Factual/Foursquare), und Namen wie Peter Norvig sind Berater (Source: commoncrawl.org). Die täglichen Abläufe stützen sich jedoch auf ein winziges festes Personal und Beiträge von Freiwilligen und Kollaborateuren.

Die Organisation wird hauptsächlich durch **Spenden und Sponsoring** finanziert, insbesondere von Cloud-Anbietern. Ab 2012 hostet Amazon Web Services (AWS) die Daten von Common Crawl kostenlos im Rahmen des AWS Public Datasets-Programms (Source: alchetron.com). Das öffentliche Datensponsoring von AWS stellt den immensen benötigten Speicherplatz (viele hundert Terabyte) zur Verfügung, ohne Common Crawl Kosten zu berechnen. Andere Cloud-Plattformen (z.B. Microsoft Azure, Google Cloud) können ebenfalls an Archiven beteiligt sein, aber AWS ist der primäre Host. Darüber hinaus haben Unternehmen wie Amazon Kleinprojekt-Wettbewerbe (z.B. 50 $ AWS-Guthaben) angeboten, um die Nutzung der Daten zu fördern (Source: commoncrawl.org). Die Stiftung erhält wahrscheinlich auch bescheidene philanthropische Spenden, obwohl **Common Crawl nie Wagniskapitalinvestitionen angenommen oder als kommerzielles Unternehmen geführt wurde**. (Es bleibt bewusst gemeinnützig, um „neutral“ und frei von Gewinnmotiven zu bleiben (Source: www.novaspivack.com) (Source: www.96layers.ai).)

Kurz gesagt, Common Crawl ist das kollaborative Produkt einiger leidenschaftlicher Technologen und des Cloud-Computing-Ökosystems. Seine relativ niedrigen Betriebskosten (da es Speichergebühren umgeht) ermöglichen es ihm, mit minimaler Finanzierung zu bestehen. Ab 2024 ist Common Crawl „der breiten Öffentlichkeit weitgehend unbekannt“, doch es wird anerkannt, dass es „eine wichtige Rolle“ in Bereichen wie der generativen KI spielt (Source: www.mozillafoundation.org). Der Bericht der Mozilla Foundation von 2024 betont, dass Common Crawl „eine kleine gemeinnützige Organisation“ mit massivem Einfluss ist (Source: www.mozillafoundation.org).

Datenerfassung: Crawling und Technologie

Common Crawl betreibt einen automatisierten Web-Crawler (genannt **CCBot**), der kontinuierlich das öffentliche Web scannt, um seinen Datensatz aufzubauen. Der Crawler basiert auf dem Open-Source-Framework Apache Nutch, das die URL-Erkennung, das Abrufen von Seiten und das Folgen von Hyperlinks übernimmt (Source: datadome.co). (Tatsächlich wechselte Common Crawl 2013 von einem benutzerdefinierten Crawler zu Apache Nutch als Kern-Crawler (Source: alchetron.com), und migrierte gleichzeitig vom älteren „ARC“-Dateiformat zum Standard-**WARC**-Format (Source: alchetron.com).) CCBot identifiziert sich im User-Agent als „CCBot/2.0“ (Source: datadome.co), obwohl es nicht ratsam ist, sich ausschließlich auf die User-Agent-Zeichenfolge zu verlassen, da Bots Identitäten fälschen können. CCBot crawlt von Amazon AWS IP-Adressen. In früheren Jahren waren die IP-Bereiche von CCBot öffentlich dokumentiert (z.B. `38.107.191.66` – `38.107.191.119`) (Source: datadome.co), aber jetzt ist der Crawler vollständig cloudbasiert.

Robots.txt und Ethik: Wie gutmütige Crawler **respektiert CCBot `robots.txt`-Regeln und Nofollow-Tags** (Source: alchetron.com), sodass er Seiten meidet, die von Website-Betreibern explizit ausgeschlossen wurden. Er konzentriert sich auf öffentlich zugängliche Inhalte (HTML-Seiten) und speichert den rohen Seiteninhalt (HTML und Text) in den Crawl-Archiven. Im Gegensatz zum Internet Archive, das Seiten zum Zweck der Archivierung und Wiedergabe (einschließlich Bilder, Skripte und clientseitiger Verhaltensweisen) erhalten möchte (Source: dallascard.github.io), konzentriert sich Common Crawl auf textuelle Inhalte und Metadaten, die für Data Mining und maschinelles Lernen nützlich sind. Insbesondere speichert oder analysiert Common Crawl *keine* Bilder, Videos, CSS oder andere statische Ressourcen im Detail – der Schwerpunkt liegt auf *rohem HTML-Text und zugehörigen Metadaten*. Dies macht den Common Crawl-Korpus direkter für NLP und Datenanalyse nutzbar, auf Kosten eines vollständigen visuellen Schnappschusses.

Crawl-Methodik: Common Crawl führt typischerweise einen **monatlichen Crawl** durch, was bedeutet, dass CCBot etwa einen Monat lang kontinuierlich Seiten abruft und die Ergebnisse dann als „Crawl-Archiv“ veröffentlicht. Dies wird ungefähr jeden Monat wiederholt. Historisch gesehen hat der Zeitplan variiert: In den frühesten Jahren gab es etwa 4 Crawls pro Jahr (Source: alchetron.com), später wurde es monatlich. Jeder monatliche Crawl beginnt mit einem riesigen Satz von Seed-URLs (anfängliche Einstiegspunkte) im öffentlichen Web und folgt Links, um neue URLs zu entdecken, wobei er unterwegs mithilfe domänenbasierter Heuristiken bereinigt, um eine breite Abdeckung zu gewährleisten. Das Ergebnis jedes Crawls ist eine Sammlung von WARC-Dateien

(komprimierte Archive abgerufener Seiten) sowie begleitende Metadaten (z.B. Tabellen von URLs, Textauszüge, Linkgraphen) (Source: alchetron.com). Um Mitte 2012 begann Common Crawl auch, Text und Metadaten, die aus jedem Crawl extrahiert wurden, zu veröffentlichen, anstatt nur rohe WARCs (Source: alchetron.com).

Umfang und Wachstum: Der Umfang des Betriebs von Common Crawl ist immens. Laut einem Interview aus dem Jahr 2023 sammelt Common Crawl *jeden Monat* **3 bis 5 Milliarden Webseiten**, was „500-mal mehr Webseiten als [die gesamte Wikipedia]" sind (Source: www.96layers.ai). Die monatlich komprimierten Daten belaufen sich auf etwa **90 Terabyte** (ungefähr 400 TB unkomprimiert) (Source: www.96layers.ai). Über mehr als ein Jahrzehnt hinweg hat Common Crawl **Hunderte von Milliarden Seiten** angesammelt. In einem Bericht (April 2024) wurde festgestellt, dass „Common Crawl in seiner 17-jährigen Geschichte mehr als *250 Milliarden Webseiten* gesammelt hat" (Source: www.96layers.ai). Die eigene Homepage (Stand Ende 2025) gibt an: *„Über 300 Milliarden Seiten aus 15 Jahren"* (Source: commoncrawl.org). (Diese Zahlen sind angesichts des kontinuierlichen Crawlings weitgehend konsistent.) Zum Vergleich: Bei seiner Einführung Anfang 2013 umfasste der erste Datensatz von Common Crawl etwa **5 Milliarden Seiten** (≈81 Terabyte) (Source: nonprofitquarterly.org) (Source: www.thekurzweillibrary.com). Mitte 2015 umfassten die archivierten Crawls über 4 jährliche Crawls hinweg etwa **1,8 Milliarden Seiten (145 TB)** (Source: alchetron.com). Heute übertrifft allein der monatliche Crawl diese früheren Gesamtzahlen.

Zusätzlich zu den Seiteninhalten veröffentlicht Common Crawl auch **Host- und Domain-Level-Linkgraphen** und andere abgeleitete Datensätze (z. B. URLs, die eine bestimmte Abfrage enthalten, oder Domain-Level-PageRank-Approximationen). Diese sind auf der *Daten*-Seite und auf GitHub verfügbar und werden regelmäßig aktualisiert. Die rohen WARC-Archive und der verarbeitete Text werden in **Amazon S3** (AWS Public Dataset) und auf Spiegelservern gehostet. Benutzer können spezifische Monats-/Jahres-Crawls per HTTP herunterladen oder Big-Data-Tools (z. B. Amazon Athena, Spark) verwenden, um die Daten direkt abzufragen. Common Crawl bietet auch Hilfswerkzeuge und Indizes (z. B. einen URL-Index) an, um die Suche nach interessanten Seiten zu erleichtern.

Insgesamt hat sich die Crawling-Technologie von Common Crawl weiterentwickelt, ist aber offen geblieben. Sie verwendet standardmäßige, bekannte Komponenten (Apache Nutch, Amazon Cloud) und Open-Source-Code für die Datenverarbeitung. Da es sich um ein gemeinnütziges Projekt handelt, nutzt es die Cloud auf kreative Weise: Es vermeidet Speicherkosten, indem es im kostenlosen AWS-Tier bleibt, und umgeht die Gebühren für die Datenübertragung (Egress), indem es die Analyse auf der AWS-Plattform fördert. Die Kerninfrastruktur von Common Crawl ist relativ einfach, aber das Ergebnis ist immens: Terabytes offener Webdaten, die als gemeinsame Ressource aggregiert und gepflegt werden (Source: www.96layers.ai) (Source: dallascard.github.io).

Datensatz und Statistiken

Der öffentliche Datensatz von Common Crawl ist eines der größten Textkorpora überhaupt, vergleichbar im Umfang mit dem Speicher großer Suchmaschinen. Wichtige Statistiken über das Korpus (Stand Mitte 2025) sind:

- **Größe des Korpus:** Über *300 Milliarden einzigartige Webseiten* (HTML-Dokumente) gesammelt (Source: commoncrawl.org). (Zum Vergleich: Dies ist Tausende Male größer als die gesamte englische Wikipedia.)
- **Zeitlicher Umfang:** Monatliche Schnappschüsse von 2008 oder 2009 bis heute (über 15 Jahre) (Source: commoncrawl.org). Jeder Schnappschuss enthält typischerweise Seiten, die in diesem Monat gecrawlt wurden. Die Sammlung wächst jedes Jahr additiv.
- **Monatliche Wachstumsrate:** Typischerweise **3–5 Milliarden Seiten pro Monat**, was etwa **90 TB komprimiert (~400 TB unkomprimiert)** pro Monat ergibt (Source: www.96layers.ai) (Source: commoncrawl.org). Über ein Jahr sind das in der Größenordnung von **30–60 Milliarden Seiten** und Hunderte von Terabytes.
- **Crawl-Frequenz:** Im Allgemeinen ein Crawl pro Monat (obwohl es anfangs weniger waren). Das Archiv ist in dem Sinne kumulativ, dass jeder Crawl ein neuer Schnappschuss ist, aber in der Praxis können Benutzer Daten aus mehreren Monaten kombinieren.
- **Datenvolumen:** Hunderte von Terabytes pro Crawl, verteilt auf WARC-Dateien, plus abgeleiteter Text und Metadaten in angrenzenden Dateien. Zum Beispiel betrug der erste Crawl von 2013 81 TB (Source: nonprofitquarterly.org), und moderne Crawls sind größer. Insgesamt belaufen sich die Archive von Common Crawl auf **mehrere Petabytes** komprimierter Daten (ein Mozilla-Bericht von 2024 nennt „mehr als 9,5 Petabytes" Common Crawl-Daten) (Source: www.mozillafoundation.org).
- **Nutzung in der Forschungsliteratur:** Über **10.000 Forschungsarbeiten** haben Common Crawl als Datenquelle zitiert (Source: commoncrawl.org) (Source: dallascard.github.io). Diese Zahl scheint sich alle paar Jahre ungefähr verdoppelt zu

haben. (Die genaue Zahl ist schwer zu überprüfen, aber die Website behauptet stolz „in über 10.000 Forschungsarbeiten zitiert“ (Source: commoncrawl.org), und unabhängige Daten zeigen, dass die Zahl 2013 viel niedriger war.)

Diese groben Zahlen verdeutlichen den immensen Umfang der Daten. Es ist bemerkenswert, dass nur wenige private Organisationen (Google, Microsoft, Amazon, Facebook) über vergleichbare Web-Scale-Crawling-Fähigkeiten verfügen – und diese Daten proprietär halten. Im Gegensatz dazu ist das Archiv von Common Crawl öffentlich auf AWS Open Data und anderen Spiegelservern gelistet, sodass **jeder** es herunterladen oder analysieren kann (Source: registry.opendata.aws).

Wichtig ist, dass Common Crawl klarstellt, dass sein Datensatz **nicht** das „gesamte Web“ ist oder garantiert vollständig ist. Die Abdeckung ist auf englischsprachige, zugängliche Webseiten ausgerichtet (über robots.txt blockierte Seiten werden ausgeschlossen, und große Plattformen wie Facebook blockieren das Crawling). Eine Mozilla-Studie von 2024 warnte ausdrücklich: *„Die unkritische Behandlung von Common Crawl als ‚Kopie des Webs‘ erklärt einen relativ kleinen Teil hauptsächlich englischer Webseiten als repräsentativ für die ganze Welt.“* (Source: www.mozillafoundation.org). In der Praxis repräsentiert Common Crawl das **„sichtbare Web“** (der Teil, der über typische HTML-Links erreichbar ist) zum jeweiligen Crawl-Datum, mit einem Schwerpunkt auf Vielfalt (es konzentriert sich nicht ausschließlich auf Top-Domains) und Aktualität.

Trotz Einschränkungen macht die schiere Breite der Daten von Common Crawl sie äußerst wertvoll. Sie **übertrifft bei Weitem** jeden statischen Datensatz, den die meisten Forscher selbst sammeln könnten. Moderne Modelle für natürliche Sprache verwenden häufig **Hunderte von Milliarden Wörtern** aus Common Crawl. Zum Beispiel wurde das Stanford GloVe Word Embedding (2014) auf **840 Milliarden Tokens** trainiert, die von Common Crawl gescrapt wurden (Source: huggingface.co). Und große LLMs nehmen routinemäßig Tausende informeller Webseiten von Common Crawl auf (wie unten beschrieben). Die Daten werden auch in der Webgraphenanalyse, der Forschung zur Informationswiederherstellung (z. B. beim Aufbau von Suchmaschinen für den ClueWeb-Datensatz (Source: commoncrawl.org) und im domänenspezifischen Mining (wie dem Extrahieren paralleler Texte für die maschinelle Übersetzung (Source: huggingface.co) verwendet.

Tabelle 1 unten fasst einige dieser Schlüsselkennzahlen und Fakten zusammen:

KENNZAHL/FAKT	WERT/BESCHREIBUNG	QUELLE
Gründungsjahr	2007 (als 501(c)(3) gemeinnützige Organisation im Jahr 2007 gegründet)	[9†L0-L4], [7†L19-L24]
Gründer und Vorsitzender	Gil Elbaz (Technologe, Mitbegründer von Applied Semantics/AdSense)	[47†L0-L4], [6†L144-L152]
Beirat (bemerkenswert)	Peter Norvig von Google, Joi Ito vom MIT, Nova Spivack, Carl Malamud	[30†L36-L38], [47†L19-L24], [45†L10-L18]
Organisationstyp	501(c)(3) gemeinnützige Organisation (Kalifornien)	[9†L0-L4], [7†L19-L24]
Alter/Zeitspanne des Datensatzes	2008/2009 – heute (über 15 Jahre monatlicher Webseiten)	[9†L10-L17], [2†L20-L24]
Gesamt gesammelte Seiten	**~300+ Milliarden** Webseiten (kumulativ)	[9†L10-L17], [2†L20-L24]
Monatliches Wachstum (Seiten)	~3–5 Milliarden neue Seiten pro Monat hinzugefügt (Durchschnitt)	[2†L20-L24], [9†L14-L17]
Monatliche Datengröße	~90 Terabyte komprimiert (~400 TB unkomprimiert) pro monatlichem Crawl	[2†L20-L24]
Einschlusskriterien	Öffentliche HTML-Seiten (unter Beachtung von robots.txt); Fokus auf Rohdaten (keine Bilder/Videos).	[52†L22-L31], [19†L28-L31]
Bemerkenswerte Projektverwendungen	KI/ML-Training (GPT-3, PaLM usw.), Word Embeddings (GloVe 840B Tokens), Forschungskorpora (C4, The Pile), Suchmaschinen	[60†L23-L30], [61†L32-L39], [52†L49-L57]
Forschungszitate (ca.)	**>10.000** veröffentlichte Arbeiten, die Common Crawl zitieren	[9†L12-L17], [52†L34-L40]
Amazon-gehosteter Datensatz	Gehostet über AWS Open Data (kostenlos für Benutzer über S3/Athena/AWS)	[19†L33-L39], [25†L12-L19] (AWS registry)
Größte LLM-Abdeckung	~80–85 % der Trainings-Tokens von GPT-3 stammen von Common Crawl (Source: www.mozillafoundation.org); ~64 % der untersuchten LLMs (2019–2023) verwenden CC (Source: www.mozillafoundation.org).	

(Tabelle 1: Wichtige Fakten und Statistiken zu Common Crawl, mit Quellenangaben.)

Geschichte und Entwicklung

Die Entwicklung von Common Crawl lässt sich chronologisch anhand mehrerer wichtiger Meilensteine betrachten:

- **2007 – Projektbeginn:** Gil Elbaz „trat an mich heran mit einer ehrgeizigen Vision – er wollte einen offenen, gemeinnützigen Crawl des Webs schaffen“ (Source: www.novaspivack.com). Im Jahr 2007 gründete er offiziell die Common Crawl Foundation. Zu den frühen Kollaborateuren gehörten Nova Spivack und Carl Malamud, die Vorstandsmitglieder wurden (Source: www.novaspivack.com). In dieser Phase arbeiteten nur wenige Personen daran (Elbaz selbst, Ahad Rana als leitender Ingenieur, einige Freiwillige). Spivack berichtet: *„Gil und der leitende Ingenieur, Ahad Rana, machten sich dann an die Arbeit, das Ding tatsächlich zu bauen.“* (Source: www.novaspivack.com). Ziel war es, „den ersten wirklich offenen, gemeinnützigen Suchindex des Webs mit 5 Milliarden Seiten“ zu schaffen (Source: www.novaspivack.com). (Tatsächlich enthielten die ersten um 2013 veröffentlichten Crawl-Daten etwa 5 Milliarden Seiten, 81 TB, wie vom MIT Tech Review berichtet (Source: nonprofitquarterly.org) (Source: www.thekurzweillibrary.com).)

- **2008–2011 – Frühe Crawls:** Nach der Gründung begann Common Crawl mit monatlichen (~vierteljährlichen) Crawls eines Teils des Webs. In diesen Jahren waren die Datenmengen kleiner; frühe Blogbeiträge deuten auf nur wenige Terabytes pro Crawl hin. Der Schwerpunkt lag auf dem Aufbau der Pipeline (Nutch-basierter Crawler, WARC-Archive, einfache Hadoop-Prozesse zur Textextraktion). Ursprünglich schrieb das Team eigenen Code, aber 2013 kündigten sie den Wechsel zu Apache Nutch und die Einführung des WARC-Dateiformats für alle Crawl-Daten an (Source: alchetron.com). Die Nutzung von Amazon S3 zur Speicherung begann wahrscheinlich in dieser Ära.

- **2012 – Partnerschaft mit Amazon AWS:** Ein wichtiger Wendepunkt ereignete sich 2012, als **Amazon Web Services** Common Crawl in sein Public Datasets-Programm aufnahm (Source: alchetron.com). AWS erklärte sich bereit, die Crawl-Archive kostenlos in seiner Cloud zu hosten. Dies war entscheidend – es ermöglichte Common Crawl, von Gigabytes auf Petabytes zu skalieren, ohne Speicherkosten tragen zu müssen. (Parallel dazu arbeiteten AWS und Common Crawl später bei Wettbewerben zusammen; z. B. bot AWS den Wettbewerbsteilnehmern 50 $ Guthaben für die Nutzung der Daten an (Source: commoncrawl.org).) Ebenfalls Ende 2012 spendete das Suchmaschinenunternehmen *Blekko* Metadaten aus seinen eigenen Crawls (Februar–Oktober 2012) an Common Crawl (Source: alchetron.com). Blekkos Protokolle halfen, die Crawl-Abdeckung zu verbessern und unerwünschte Seiten (Spam, Pornografie, SEO-Manipulationen) zu reduzieren (Source: alchetron.com).

- **2013 – Offizieller Start und Anerkennung:** Anfang 2013 erregte die erste große öffentliche Veröffentlichung von Common Crawl (der „5-Milliarden-Seiten-Index“) Medienaufmerksamkeit. Das MIT Technology Review (über Ray Kurzweils Blog) veröffentlichte im Januar 2013 einen Artikel mit dem Titel „Eine kostenlose Datenbank des gesamten Webs könnte das nächste Google hervorbringen“ (Source: www.thekurzweillibrary.com). Der Artikel betonte, dass *„Common Crawl über fünf Milliarden Webseiten kostenlos zur Verfügung stellt, damit Forscher und Unternehmer Dinge ausprobieren können, die sonst nur denen mit Zugang zu Googles Ressourcen möglich wären.“* (Source: www.thekurzweillibrary.com). Zu diesem Zeitpunkt waren Peter Norvig und Joi Ito dem Beirat beigetreten (Source: www.thekurzweillibrary.com). Die eigene Website und das Dashboard von Common Crawl wurden gestartet, die das jahrzehntelange Datenarchiv bewarben und die ersten Forschungsnutzer gewannen.

- **2014–2019 – Datenerweiterung und Ökosystemwachstum:** Mitte der 2010er Jahre setzte Common Crawl die monatlichen Crawls fort, und der kumulative Datensatz wuchs rapide. Jedes Jahr wurden mehr Forschung und Entwicklung auf diesen Daten aufgebaut. Wichtige Ereignisse sind:

  - **2014–2015:** Extraktion strukturierter Daten: Common Crawl begann, Text und Metadaten aus den Rohseiten zu extrahieren und diese zusammen mit den WARC-Dateien zu veröffentlichen. Daten für Sprachen wie Spanisch, Deutsch usw. wurden verfügbar gemacht. Die Community entwickelte auch Tools zur direkten Abfrage der Daten, wie z. B. Recipes und Index (über AWS Athena).
  - **2016:** Einführung von **CCBot v2.0** mit aktualisiertem User-Agent (Source: datadome.co) und Verbesserungen bei der Einhaltung von robots.txt. Die Rolle von Common Crawl in der Forschung wurde gefestigt, da NLP-Aufgaben wie GloVe (84 GB) CC-Daten verwendeten (Source: huggingface.co).

- **2017–2019:** Der Datensatz überschritt zig Milliarden Seiten. In dieser Zeit initiierte Europa den *Norvig Web Data Science Award* (unterstützt von Common Crawl und SURFSara), der die akademische Nutzung der Daten förderte. Auch das Kernentwicklungsteam blieb klein; in Interviews wurde erwähnt, dass es um 2017 nur drei Mitarbeiter hatte (Source: www.96layers.ai). Bis 2019 wurde Common Crawl als wichtige Quelle für das Training neuronaler Modelle anerkannt, blieb aber in der breiten Öffentlichkeit noch weitgehend unbemerkt.

- **2020–2022 – KI-Boom:** Der KI-Boom der COVID-Ära rückte Common Crawl ins Rampenlicht. OpenAIs GPT-3 (Mitte 2020 veröffentlicht) nutzte Common Crawl als primäre Datenquelle. Forschungsteams hinter Modellen wie Grover (Zellers et al., 2019) trainierten explizit mit CC für die Generierung von Fake News (Source: dallascard.github.io). Metas RoBERTa (2019) und Googles T5 griffen ebenfalls auf von CC abgeleitete Korpora zurück. Im Jahr 2020 wurden die Daten von Common Crawl in große Forschungsdatensätze wie „C4“ (für T5 verwendet) und „The Pile“ (ein 800 GB großes englisches Korpus) integriert – beide erkennen CC öffentlich als Hauptbestandteil an (Source: dallascard.github.io). Die Öffentlichkeit begann von „Billionen von Tokens“ zu hören, die für KI aus dem Web gescrapt wurden, und Common Crawl wurde als Schlüsselquelle identifiziert. Common Crawl selbst blieb jedoch klein; es wurde berichtet, dass die Organisation zum Zeitpunkt der Einführung von GPT-3 möglicherweise nur einen einzigen Mitarbeiter hatte, der daran arbeitete (Source: www.96layers.ai).
- **2023–2025 – Aktuelle Ära und öffentliche Anerkennung:** In den Jahren 2023 und 2024 erfuhr Common Crawl aufgrund zweier Faktoren einen Anstieg der öffentlichen Aufmerksamkeit: (a) der Aufstieg der generativen KI, für die die offenen Daten von CC unerlässlich sind; und (b) rechtliche Kontroversen um urheberrechtlich geschütztes Material in Trainingsdaten. Anfang 2024 veröffentlichte die Mozilla Foundation einen ausführlichen Bericht (basierend auf Interviews mit Common Crawl-Mitarbeitern) mit dem Titel *„Trainingsdaten zum Preis eines Sandwiches: Der Einfluss von Common Crawl auf generative KI.“* (Source: www.mozillafoundation.org). Dieser Bericht enthüllte aktuelle Statistiken (9,5 PB Daten, 84 % der GPT-3-Tokens von CC) und lieferte aktualisierte Einblicke in die Organisation. Etwa zur gleichen Zeit brachte ein bemerkenswerter Rechtsstreit (New York Times gegen OpenAI/Microsoft) Common Crawl in die Schlagzeilen, da NYT-Inhalte in CC gescrapt und somit unbeabsichtigt in GPT-3 verwendet wurden (Source: www.mozillafoundation.org). Das Common Crawl-Team kündigte auch neue Dienste an (z. B. das Hosten eines abfragbaren *Common Crawl Index* (Source: commoncrawl.org) und erweiterte das Community-Engagement (Artikel, Tutorials, Hackathons).

Im Laufe seiner Geschichte ist Common Crawl seiner ursprünglichen Mission des **offenen Zugangs** treu geblieben. Es hat sich nie zu einer kommerziellen Suchmaschine oder einem Datenanbieter entwickelt. Stattdessen konzentrierte es sich auf den Aufbau einer robusten, skalierbaren Pipeline und einer Community rund um offene Daten. Die Projektleitung betont regelmäßig, dass *„die Bereitstellung von KI-Trainingsdaten nie der Hauptzweck von Common Crawl war“* und dass sie stets eine breite Nutzerbasis (KI-Forscher sind nur eine Gruppe) willkommen geheißen hat (Source: www.96layers.ai). Dennoch hat, wie wir noch erörtern werden, das Aufkommen generativer KI Common Crawl einflussreicher denn je gemacht – sowohl im Guten (Ermöglichung von Forschung) als auch im Kontroversen (Urheberrechts- und Voreingenommenheitsbedenken).

Technische Details der Common Crawl Daten

Datenformate und Zugang

Jeder Common Crawl-Durchlauf erzeugt eine Reihe von Dateien im **WARC**-Format (Web ARChive), das Sequenzen von HTTP-Antworten (die abgerufenen Webseiten) mit Metadaten bündelt. Diese WARC-Dateien sind die rohe Crawl-Ausgabe, typischerweise benannt nach Datum und Crawl-Identifikator. Zusätzlich zu den WARC-Dateien veröffentlicht Common Crawl eine Vielzahl begleitender Dateien:

- **Extrahierter Text (WAT-Dateien):** Für jede WARC-Datei enthält eine entsprechende „WAT“-Datei geparste Metadaten (z. B. HTTP-Header, Links, JSON-Metadaten).
- **Extrahierter Text (WET-Dateien):** Eine „WET“-Datei streamt den aus jeder HTML-Seite extrahierten Klartext (im Wesentlichen den bereinigten Textinhalt). Diese ermöglichen es Benutzern, Text schnell zu analysieren, ohne selbst das HTML parsen zu müssen.
- **URL-Index (CDX):** Ein CSV/JSON-Index aller abgerufenen URLs und ihrer Offsets in WARC-Dateien, nützlich für die Abfrage spezifischer Websites oder Seiten.
- **Web-Graphen:** Graphdaten, die Seiten oder Domains verknüpfen (z. B. Host-zu-Host-Link-Graphen). Diese werden regelmäßig (z. B. jährlich) bereitgestellt, um die Konnektivität zu untersuchen.
- **Domain-Tabellen:** Aggregierte Dateien, die alle gecrawlten Domains und Seitenzahlen auflisten.

Alle diese Dateien werden in **AWS S3 Buckets** gespeichert (und anderswo gespiegelt). Common Crawl fördert die Nutzung von In-Cloud-Analysen (z. B. Amazon Athena oder EMR), um die Daten im großen Maßstab abzufragen. Zum Beispiel ermöglicht Amazon Athena SQL-Abfragen über den Index aller URLs oder sogar den WARC-Inhalt, wenn dieser richtig strukturiert ist. Die Kosten für die

Ausführung solcher Abfragen sind gering (und manchmal durch Guthaben gedeckt), was es Forschungsteams praktisch macht, Datensätze aus Common Crawl zu extrahieren, ohne Terabytes auf ihre lokalen Server kopieren zu müssen.

Common Crawl selbst bietet einige Entwicklertools und Dokumentationen (z. B. das Projekt „Index to WARC Files and URLs“ (Source: registry.opendata.aws). Es gibt aber auch ein lebendiges externes Ökosystem: Zahlreiche GitHub-Projekte und Tutorials (z. B. CC-pyspark, commoncrawljob) helfen neuen Benutzern beim Einstieg. Die öffentliche Mailingliste von Common Crawl sowie die Slack/Discord-Communities sind aktiv mit Tipps und geteiltem Code.

CCBot (Common Crawl Crawler)

Der **Web-Crawler** selbst, genannt **CCBot**, läuft während jedes monatlichen Crawls kontinuierlich. Er funktioniert ungefähr so: Ein Master-Scheduler verteilt Crawler-Instanzen (auf AWS EC2), die Seiten parallel abrufen und der Liste der zu besuchenden URLs folgen. Neue URLs werden der Warteschlange hinzugefügt, sobald Links entdeckt werden. Der Crawler nutzt die Standardfunktionen von Nutch: Respekt vor `robots.txt`, automatische Drosselung pro Domain und Deduplizierungslogik, um ein endloses Crawlen desselben Inhalts zu vermeiden (z. B. Entfernen von Session-Parametern).

CCBot identifiziert sich mit einem User-Agent-String, aber Common Crawl empfiehlt Webmastern, nicht ausschließlich danach zu whitelisten, da betrügerische Crawler diesen fälschen können (Source: datadome.co). (Stattdessen können Website-Betreiber bekannte AWS-IP-Bereiche verwenden, um CCBot-Traffic zu identifizieren.) Obwohl CCBot ein legitimer Nutzer ist, stammen seine IP-Adressen aus dynamischen AWS-Pools, sodass einige Websites ihn unbeabsichtigt blockieren oder drosseln. Common Crawl bemüht sich, ein „höflicher“ Crawler zu sein. Zum Beispiel wechselt es IP-Bereiche, zieht sich von überlasteten Websites zurück und erlaubt einige Crawl-Fehler. Serveradministratoren, die Community-Normen respektieren möchten, können CCBot explizit zulassen, indem sie ihre `robots.txt` anpassen (Common Crawl bietet Dokumentation dazu).

Im Laufe der Zeit wurde CCBot auf Effizienz hin verfeinert. Die aktuelle Architektur (Stand 2025) verwendet ein verteiltes, fehlertolerantes System auf AWS, koordiniert vom Kernteam (geleitet von einem „Crawl Engineer“). Der Crawl vom Mai 2025 umfasste beispielsweise **2,47 Milliarden Seiten** (siehe Twitter-Gipfelbericht (Source: commoncrawl.org). Alles in allem hat sich das System als **skalierbar** erwiesen: Common Crawl bemerkt stolz, dass sein Crawl inzwischen *„gigantisch“* ist und weit über die Kapazität jedes akademischen Forschers hinausgeht, ihn zu duplizieren (Source: nonprofitquarterly.org).

Datenverarbeitungspipeline

Roh gecrawlte Seiten durchlaufen vor der Veröffentlichung eine Verarbeitungspipeline. Wichtige Schritte umfassen:

- **Link-Extraktion:** Alle Hyperlinks auf jeder Seite identifizieren, um sie der Crawl-Frontier hinzuzufügen. Link-Graphen (auf Domain- und Host-Ebene) für die Analyse erstellen.
- **Inhalts-Deduplizierung:** Identische oder nahezu identische Seiten herausfiltern, um Verschwendung und Verzerrungen zu reduzieren. Common Crawl wendet eine aggressive Deduplizierung auf Dokument- und Seitenebene an, damit archivierte Daten minimale Redundanz aufweisen.
- **Text-Extraktion:** HTML/CSS entfernen und Textinhalte extrahieren, die in den „WET“-Dateien gespeichert werden. Dies beinhaltet die Spracherkennung (Common Crawl konzentriert sich typischerweise auf englischen Text, erfasst aber auch andere Sprachen).
- **HTTP-Metadaten:** Die Antwort-Header, den Inhaltstyp und Serverinformationen für jeden Abruf aufzeichnen (in den WAT-Dateien).
- **Fehlerbehandlung:** Alle Abruffehler oder Timeouts in einer „Errata“-Datei aufzeichnen. Common Crawl führt **ein Errata-Log** mit URLs oder Domains, die konsistent fehlschlagen, um zukünftige Crawls zu verbessern.

Das Endergebnis ist ein reichhaltiges Datenprodukt: Für jeden beliebigen Monat kann ein Benutzer nicht nur die rohen HTML-Blobs abrufen, sondern auch ein paralleles Satzkorpus (den WET-Text) und die gesamte Hyperlink-Struktur. Der Pipeline-Code ist Open Source, und Verbesserungen (z. B. besseres HTML-Parsing, JavaScript-Verarbeitung) werden regelmäßig integriert.

(Im Februar 2023 kündigte Common Crawl in seinem Blog an, mit dem *Pre-Rendering* von Seiten, die JavaScript erfordern, experimentieren zu wollen – doch Ende 2025 bleibt das Hauptkorpus HTML-zentriert.)

Datensatzmerkmale

- **Sprachverteilung:** Die Menüs von Common Crawl zeigen, dass der Datensatz mehrsprachig ist, aber stark zum Englischen tendiert. Laut Mozillas Bericht ist der Crawl **„primär englisch“**, wobei die regionale Abdeckung variiert. Zum Beispiel wurden Datensätze von 50 Millionen deutschen Nachrichtenartikeln (Source: commoncrawl.org) und andere sprachspezifische Korpora aus CC abgeleitet, aber der Roh-Crawl enthält weitaus mehr englische Inhalte.
- **Website-Vielfalt:** Common Crawl versucht, Breite und Tiefe auszugleichen. Es umfasst große Websites (Nachrichten, E-Commerce, Blogs) sowie Long-Tail-Websites. Es zielt jedoch nicht auf das „Deep Web“ oder passwortgeschützte Seiten ab. Es kann auch keine Websites crawlen, die Bots nicht zulassen oder Anmeldungen erfordern.
- **Zeitliche Schnappschüsse:** Jeder monatliche Crawl ist mit einem Zeitstempel versehen. Folglich können Common Crawl-Archive verwendet werden, um die Entwicklung des Webs zu untersuchen (z. B. wie sich eine Seite oder Domain im Laufe der Zeit ändert). Common Crawl ist jedoch *kein* kontinuierliches Archiv wie die Wayback Machine – es bewahrt nicht jede Version einer Seite täglich auf; hauptsächlich bietet es einen „Snapshot“ pro URL pro Monat (es sei denn, die Seite ändert sich und wird später erneut gecrawlt).

Zusammengenommen sind die Daten von Common Crawl extrem groß und ziemlich repräsentativ für das öffentliche Web (vorbehaltlich Bots und Zugang). Es ist *das* größte öffentlich zugängliche Webarchiv für Forschungszwecke, das Volumen mit Zugänglichkeit verbindet.

Anwendungsfälle und Auswirkungen

Der offene Datensatz von Common Crawl hat eine Vielzahl von Anwendungen ermöglicht. Wir gliedern seine Nutzung in mehrere breite Kategorien:

1. KI und Maschinelles Lernen (LLMs, Embeddings usw.)

Common Crawl ist zur **Eckpfeiler-Datenquelle für die großskalige Verarbeitung natürlicher Sprache und KI** geworden. Praktisch jedes moderne Sprachmodell hat auf diese Daten zurückgegriffen. Zum Beispiel:

- **GPT-3 und ChatGPT:** Als OpenAI GPT-3 (das ChatGPT zugrunde liegt) trainierte, stammte der Großteil seiner Trainings-Tokens von Common Crawl. OpenAIs veröffentlichtes GPT-3-Paper zeigt, dass *„die größte Menge an Trainingsdaten von Common Crawl stammt“* (Source: datadome.co). Eine Mozilla-Analyse bestätigt dies: Sie ergab, dass *über 80 % der GPT-3-Tokens* von Common Crawl stammten (Source: www.mozillafoundation.org). (GPUs trainieren typischerweise mit mehreren Korpora; für GPT-3 waren die anderen Quellen WebText2, Bücher und Wikipedia. Aber Common Crawl war der größte Anteil.) Da GPT-3 direkt in Chatbots und KI-Assistenten einfließt, „spricht“ der Inhalt von Common Crawl (gut oder schlecht) im Wesentlichen über KI zu den Endnutzern.
- **Andere große Sprachmodelle:** Viele andere bemerkenswerte LLMs wurden auf CC-Daten aufgebaut:
  - Googles T5 und BERT-basierte Modelle integrierten Teilmengen von Common Crawl.
  - Facebooks RoBERTa wurde 2019 mit einer Mischung aus CC- und Nachrichtendaten trainiert.
  - Open-Source-Modelle wie EleutherAIs GPT-NeoX und kleinere Modelle wie GPT-2 nutzten CC.
  - Das **Grover**-Modell (2019) von Zellers *et al.* – ein Modell zur Generierung und Erkennung von Fake News – verwendete explizit Common Crawl für Webtext (Source: dallascard.github.io).
  - In jüngerer Zeit verwenden die meisten neuen Modelle (Bellatrix, LLaMA usw.) Pipelines wie The Pile oder RefinedWeb, die wiederum aus Common Crawl-Snapshots stammen (Source: dallascard.github.io). Tatsächlich werden Common Crawl-Snapshots in abgeleiteten Datensätzen (z. B. C4, Colossal Clean Crawls) neu verpackt, die große Trainings-Workloads speisen.
  - Eine Umfrage unter 47 verschiedenen LLMs (2019–2023) ergab, dass *„mindestens 64 %“* von ihnen mit Common Crawl-Daten trainiert wurden (Source: www.mozillafoundation.org). Dies umfasst neuere Modelle wie ChatGPT-4 (über GPT-4), Metas LLaMA, Mistral, Claude 2 usw. (Einige Modelle verwenden möglicherweise auch proprietäre oder gemischte Daten, aber CC bleibt eine Hauptstütze.)

- **Wort-Embeddings und NLP-Tools:** Der Datensatz hat grundlegende NLP-Ressourcen ermöglicht. Die klassischen **GloVe-Embeddings** (840 Mrd. Tokens, Englisch) und **FastText-Embeddings** (600 Mrd. Tokens) werden beide mit CC-Text trainiert (Source: huggingface.co). Open-Source-Korpora wie Colossal Clean Crawls (C4) und von Common Crawl abgeleitete mehrsprachige Datensätze treiben Übersetzungsmodelle und Summarizer an. Forschung in den Bereichen Topic Modeling, Sentiment-Analyse, Informationsabruf und mehr nutzt CC oft als Rohtextquelle. Zum Beispiel erstellte eine Studie aus dem Jahr 2019 ein bilinguales Parallelkorpus aus CC für die maschinelle Übersetzung (Source: huggingface.co).

- **Chatbots und KI-Assistenten:** Über das Offline-Modelltraining hinaus führen einige Dienste Echtzeit-Crawling von CC durch, um KI zu unterstützen. Zum Beispiel nehmen **DeepSeek** und einige „KI-gesteuerte“ Suchplattformen CC-Seiten auf, um ihre Antworten zu liefern. Viele KI-Bots verlassen sich auch auf CC, um Antworten zu überprüfen oder zu erweitern, da es ein praktischer Index für das öffentliche Web ist.

- **Daten für Vision- und multimodale Modelle:** Während Common Crawl hauptsächlich Text enthält, beinhaltet es auch URLs von Bildern (und gelegentlich Bildmetadaten). Unternehmen wie **TinEye** nutzen den Bild-URL-Index von CC, um Reverse-Image-Suchdienste aufzubauen (Source: nonprofitquarterly.org). (TinEye nutzte Common Crawl explizit, um Bilder zu finden, die einem Abfragebild ähneln.) Einige KI-Vision-Modelle verwenden CC-ausgerichtete Textbeschriftungen oder Alt-Texte in CC-Daten, um sie mit Bildern zu koppeln.

Zusammenfassend lässt sich sagen, dass **KI-Forscher und Unternehmen Common Crawl stark** als kostenlose Datenquelle **nutzen**. Seine Allgegenwart im Modelltraining hat sowohl Chancen (Fortschritt der KI) als auch Bedenken (Voreingenommenheit, Urheberrecht) aufgeworfen – mehr dazu weiter unten.

2. Akademische und wissenschaftliche Forschung

Das Common Crawl-Korpus wird in der akademischen Forschung disziplinübergreifend häufig zitiert:

- **Natürliche Sprache und Webwissenschaft:** Forscher analysieren Sprachgebrauch und -muster. Zum Beispiel wurde CC verwendet, um die Hyperlink-Struktur zu untersuchen (wer auf wen im Web verlinkt), Nachrichten geografisch zu lokalisieren (ein Datensatz von 50 Millionen deutschen Nachrichtenartikeln wurde aus CC erstellt (Source: commoncrawl.org) und die Lesbarkeit oder gängige Phrasen im Web zu analysieren. Arbeiten an **Web-Graphen** (Graphentheorie angewendet auf Domains) nutzen oft die Link-Graph-Daten von CC (Source: commoncrawl.org).

- **Data Mining und Big Data Analyse:** Der Datensatz ist ein Beispiel für „Big Open Data“. Forscher testen großskalige Text-Mining-Algorithmen (Clustering, Ausreißererkennung, Themenanalyse) auf CC. Die Möglichkeit, auf Petabytes realer Daten zuzugreifen, hat vergleichende Studien von Textverarbeitungspipelines ermöglicht.

- **Information Retrieval (IR) Studien:** Common Crawl wird zum Aufbau experimenteller Suchmaschinen verwendet. Zum Beispiel ist **Elastic ChatNoir** an der Bauhaus-Universität Weimar für die Suche in den ClueWeb- und Common Crawl-Archiven konzipiert (Source: commoncrawl.org). IR-Forscher bewerten auch Ranking-Algorithmen auf CC-Teilmengen oder verwenden CC als Referenz für Webinhaltsseiten. Das Common Crawl-Team selbst bietet eine „Simple Speedy Search“ (CCSS) API für schnelle Stichwortsuchen über den Index an.

- **Cybersicherheit und Missbrauchsmessung:** Die große Skalierbarkeit von CC ermöglicht das Scannen nach bösartigen Mustern. Zum Beispiel scannte das Paper „Lurking Malice in the Cloud“ (ACM 2016) alle CC-Seiten, um eingebettete Skripte zu finden, die mit bekannten Malware-Domains verknüpft sind (Source: huggingface.co). Forscher haben CC verwendet, um die Verbreitung von (unsicheren) HTTP-Headern, veralteten Bibliotheken oder Cryptojacking-Skripten auf beliebten Websites zu quantifizieren.

- **Wirtschafts- und Sozialwissenschaften:** Sozialwissenschaftler nutzen CC als Proxy für den öffentlichen Diskurs. Zum Beispiel nutzte eine Studie CC zur Analyse von Inhaltsmoderation und Zensur; die Forschung „Banned Books“ des *Citizen Lab* analysierte über CC gescrapte Amazon-Produktseiten, um Zensurrichtlinien zu erkennen (Source: commoncrawl.org). Weitere Anwendungsfälle umfassen die Verfolgung von Gesundheitsdesinformationen, die Analyse politischer Propaganda oder die Untersuchung der Verbreitung von Inhalten in mehreren Sprachen im offenen Web.

- **Zitationsindizes und Wissenschaftskartierung:** Die Verfügbarkeit von Milliarden wissenschaftlicher Zitationen, die aus CC-Texten gewonnen wurden, hat sogar Meta-Forschung ermöglicht. Zum Beispiel die Wiederholung von Zitationsanalysen und die Konstruktion von Wissensgraphen in kolossalem Maßstab.

Bemerkenswerterweise hebt die Common Crawl-Website selbst viele Forschungsarbeiten hervor: Sie kuratiert Links zu veröffentlichten Arbeiten, die CC-Daten nutzen (Source: commoncrawl.org). Die Zitationen umfassen NeurIPS/ICLR für NLP, WWW/WWW-Konferenzen für Webanalyse sowie Fachzeitschriften aus den Bereichen KI, Informationswissenschaft und Computational Social Science.

3. Kommerzielle und industrielle Anwendungen

Über die Wissenschaft hinaus haben zahlreiche Unternehmen und Startups Produkte auf Basis von Common Crawl-Daten entwickelt. Einige bemerkenswerte Beispiele:

- **Bildersuche – TinEye:** Wie bereits erwähnt, nutzt **TinEye** (von Idée Inc.) Common Crawl zur Indexierung von Bildern. Wenn ein Nutzer ein Bild einreicht, hasht TinEye es und durchsucht die aus CC gewonnenen Bilddaten, um ähnliche zu finden (Source: nonprofitquarterly.org). CC stellte eine große, kostenlose Quelle für Bilder und deren URLs bereit, was TinEye ermöglichte, ein tragfähiges Geschäft zu starten, ohne das Web selbst durchsuchen zu müssen.
- **Wirkungsanalyse – Lucky Oyster:** Lucky Oyster Labs (von Rendever übernommen) nutzte Common Crawl für Social Listening und Trendanalysen. Sie entwickelten Tools auf CC, um „zu verstehen, was Menschen im Web diskutieren“, als eine Art Insight-Engine (Source: nonprofitquarterly.org). (Der NPQ-Artikel erwähnt Lucky Oyster als Startup, das CC nutzt, obwohl Details jetzt spärlich sind.)
- **Search-as-a-Service – Fall Crate.IO:** Einige Unternehmen entwickelten Konnektoren und Engines, um CC-Daten abzufragen. Zum Beispiel veröffentlichte Crate.IO einen Blogbeitrag über das „Importieren aus benutzerdefinierten Datenquellen“ mithilfe eines Plugins, der zeigte, wie CC-Archive in ihre SQL-Datenbank eingespeist werden können (Source: commoncrawl.org). Ebenso sind „CommonCrawlJob“ und „CommonCrawlScalaTools“ GitHub-Projekte, die beim Laden von CC-Daten in Big-Data-Systeme helfen. Dies sind hauptsächlich Proof-of-Concept- oder Entwickler-Tools.
- **Startup-Suchmaschinen:** Mindestens ein Gründerteam (Elastic ChatNoir (Source: commoncrawl.org) entwickelte ein Suchmaschinen-Frontend speziell für Common Crawl-Klone des ClueWeb-Datensatzes. Ein weiteres, die Open-Web-Snapshots von Carrot Search, hat mit CC experimentiert. Es besteht Interesse daran, gemeinnützige oder alternative Suchmaschinen zu erstellen, die CC als Daten-Backend nutzen – wodurch die Notwendigkeit entfällt, das Web selbst zu crawlen.
- **Marketing und SEO:** Einige SEO-Analysefirmen nutzen CC, um den Website-Zugriff oder die Konkurrenzanalyse abzuschätzen. Obwohl die meisten kommerziellen SEO-Produkte auf proprietären Crawlern basieren, bietet CC einen kostenlosen Datenpool, um globale Seitenanzahlen oder Inhaltstrends zu messen. Zum Beispiel könnten die Codezeilen für SEO-Tools wie Majestic oder Ahrefs CC-Daten für die Backlink-Analyse integrieren, obwohl Details normalerweise proprietär sind.
- **Werbung und Business Intelligence:** Datenunternehmen (einschließlich Factual, das von Gil Elbaz gegründete Unternehmen) haben CC-Daten integriert, um Geschäftsdatensätze anzureichern. Zum Beispiel können Domain-Zählungen, Website-Aktualität und Inhaltsklassifizierung aus CC gewonnen werden, um Ad-Targeting- oder B2B-Marketing-Tools zu speisen. Aufgrund der automatisierten Natur der Daten müssen CC-basierte Erkenntnisse jedoch für die kommerzielle Nutzung sorgfältig validiert werden.

Tabelle 2 (unten) fasst einige illustrative Anwendungsfälle und Projekte zusammen, die Common Crawl-Daten nutzen:

NUTZER/PROJEKT	ANWENDUNGSFALL	QUELLE / ANMERKUNGEN
TinEye	Reverse Bildersuche (ähnliche Bilder durch Crawling finden)	Nutzt von CC gecrawlte Bilder (Source: nonprofitquarterly.org). (IDée Inc.)
Lucky Oyster	Analyse sozialer/kultureller Trends	Startup nutzt CC zur Analyse von Web-Content-Trends (Source: nonprofitquarterly.org).
GloVe (Stanford)	Wortvektor-Embeddings (840 Mrd. Tokens von CC)	CC stellte Text für das GloVe-Modell bereit (Source: huggingface.co).
GPT-3/ChatGPT	Trainingsdaten für großes Sprachmodell (~80% Tokens von CC)	Mozilla-Bericht: *„Über 80% der GPT-3-Tokens stammten von Common Crawl.“* (Source: www.mozillafoundation.org).
Sprachmodelle	Training/Fine-Tuning (RoBERTa, T5, LLaMA, etc.)	LLMs (2019–2023) verwenden oft CC-basierte Korpora (Source: dallascard.github.io) (Source: www.mozillafoundation.org).
Suchmaschinen	Aufbau alternativer Suchindizes (z.B. ChatNoir)	Elastic ChatNoir: Suche in CC-Daten (Source: commoncrawl.org). (Bauhaus-Weimar)
NLP-Forschung	Statistische Analyse von Webtexten (Themenmodelle, Zusammenfassung)	Dutzende akademische Arbeiten in verschiedenen NLP-Bereichen zitieren CC.
Web-Metriken	Studien zu Zensur/Meinungsfreiheit (z.B. Amazon-Zensur)	Citizen Lab „Banned Books“ nutzte CC (Source: commoncrawl.org); weitere Web-Science-Arbeiten.

(Tabelle 2: Ausgewählte Beispiele, wie Common Crawl-Daten in der Praxis genutzt werden, mit Zitationen.)

Zusätzlich zu diesen Beispielen listet die eigene Website von Common Crawl **zahlreiche Projekte** auf: offene Datensätze (WikiSQL aus Webtabellen), cloudbasierte Such-Experimente, Elasticsearch-Tutorials und akademische Kurse, die alle auf CC-Daten basieren (Source: commoncrawl.org). Anekdotisch hat Gil Elbaz kommentiert, dass **„wenn man nicht Google oder OpenAI oder Microsoft ist, fast jeder auf Common Crawl angewiesen ist“** für große Datenmengen (Source: www.96layers.ai). Dies unterstreicht, wie allgegenwärtig CC für jede Organisation geworden ist, die keinen eigenen Web-Crawler im Google-Maßstab einsetzen kann.

Fallstudien

Um die Auswirkungen von Common Crawl konkreter zu veranschaulichen, beschreiben wir zwei detaillierte Fallstudien: eine zum Thema KI/Modelltraining und eine zur offenen Suche.

Fallstudie: GPT-3 und die LLM-Revolution

Als ein prominentes Beispiel betrachten wir OpenAI’s GPT-3 (2020) und seine Schwestermodelle. Diese „Generative Pretrained Transformers“ erreichen beeindruckende Fähigkeiten in der natürlichen Sprachverarbeitung, aber ihre Leistung beruht auf riesigen Trainingsdatenmengen. Common Crawl spielte dabei eine Hauptrolle:

- **Datensatz-Zusammensetzung:** Das GPT-3-Paper (Brown et al. 2020) listet die Datenquellen auf: WebText2 (OpenAIs eigener Crawl von Reddit-verlinkten Seiten), Google Books, Wikipedia und **Common Crawl**. In Bezug auf die Rohgröße war Common Crawl bei weitem der größte. Eine spätere Analyse bestätigt, dass *„der größte Teil der Trainingsdaten von Common Crawl stammt“* (Source: datadome.co). Mozillas Bericht stellt klar, dass **über 80% aller von GPT-3 verwendeten Tokens von CC stammten** (Source: www.mozillafoundation.org).

- **Resultierendes Modell:** GPT-3-175B, mit 175 Milliarden Parametern, wurde auf 570 GB gefilterter Textdaten (rund 500 Milliarden Tokens) trainiert. Wenn 80% davon von CC stammten, bedeutet das ~456 GB CC-Text. Dieser Maßstab wäre ohne ein bestehendes Web-Korpus unmöglich. Die Verfügbarkeit von CC bedeutete, dass OpenAI zu diesem Zeitpunkt keine Ressourcen für das Crawlen des Webs selbst aufwenden musste (obwohl sie wahrscheinlich auch einige interne Webdaten hatten).

- **Professionelle Nutzung:** Als GPT-3 auf den Markt kam, wurde es schnell in Produkte integriert (z.B. Microsofts Copilot, ChatGPT von OpenAI im Jahr 2022). Diese Dienste fungieren dann als eine „KI-Schicht“ auf CC. Einige Nutzer befürchten, dass ChatGPT beim Generieren von Antworten Texte von Common Crawl-Seiten ohne Quellenangabe wiedergeben könnte. Tatsächlich stellt der Mozilla-Bericht fest, dass CC-basierte Modelle oft *voreingenommene oder urheberrechtlich geschützte Inhalte* produzieren, weil sie dazu neigen, Trainingsdaten zu memorieren.

- **Rechtliche Implikationen (NYT-Fall):** Ende 2023 verklagte The New York Times OpenAI mit der Behauptung, dass die Trainingsdaten von ChatGPT (GPT-3.5/GPT-4) Times-Inhalte unsachgemäß enthielten. Common Crawl wurde zu einem wichtigen Beweisstück, da die Artikel der Times vor dem Training des Modells in CC gescrapt worden waren und OpenAI diese CC-Snapshots nutzte. Ein Mozilla-Factsheet erklärt: *„NYT-Inhalte machten einen erheblichen Anteil der Common Crawl-Daten zu dem Zeitpunkt aus, als OpenAI ChatGPT startete, und stellten somit wahrscheinlich einen erheblichen Teil der Trainingsdaten von GPT-3 dar“* (Source: www.mozillafoundation.org). Dies verdeutlicht, wie die Offenheit von CC unbeabsichtigt zu rechtlichen Risiken führen kann, wenn urheberrechtlich geschützter Text in Modellen weiterverbreitet wird.

- **Vielfalt und Voreingenommenheit:** Da so viele LLMs auf CC basieren, verbreiten sich die in CC gelernten Direktiven weit. Wenn CC nicht genügend Inhalte aus bestimmten Sprachen oder Demografien enthält, können Modelle bei diesen Themen schlechter abschneiden. Mozillas Forschung warnt, dass *„der Datensatz von Common Crawl bewusst problematische Inhalte (Toxizität, Hassrede usw.) enthält, um die Forschung zu diesen Phänomenen zu unterstützen.“* Im Gegensatz dazu filtern viele KI-Trainingspipelines CC stark (z.B. behalten sie nur „englische, hochwertige Seiten“) (Source: www.mozillafoundation.org), was bedeutet, dass die Roh-Toxizität von CC das Modellverhalten beeinflussen kann, wenn sie nicht sorgfältig entfernt wird.

Zusammenfassend zeigt der GPT-3-Fall, dass **Common Crawl in den 2020er Jahren zum Rückgrat der generativen KI-Forschung geworden ist**. Es hat die Hürde für das Training großer Modelle dramatisch gesenkt. Die Tatsache, dass die Daten einer kleinen gemeinnützigen Organisation millionenschwere KI-Systeme antreiben, ist bemerkenswert. Es erzwingt auch eine Abrechnung: Wenn ein offener Datensatz Closed-Source-KI antreibt, wer trägt die Verantwortung für den Inhalt? Die Führung von Common Crawl betont, dass die Daten für alle Arten von Analysen (einschließlich Hassrede-Forschung) gedacht waren, nicht explizit für das Training von Milliarden-Dollar-Modellen (Source: www.96layers.ai). Die Debatte in der Gemeinschaft dreht sich nun darum, wie sichergestellt werden kann, dass CC-basierte Modelle „vertrauenswürdig“ sind (Entfernung von Voreingenommenheit, Respektierung des Urheberrechts usw.) (Source: www.mozillafoundation.org).

Fallstudie: Offene Suche über Common Crawl

Ein weiterer illustrativer Fall sind Versuche, **Suchmaschinen mithilfe von Common Crawl-Daten zu bauen**. Während Web-versierte Unternehmen wie Google oder Bing ihre eigenen Crawler entwickeln, haben einige unabhängige Gruppen die Nutzung von CC als Datenquelle für alternative Suchdienste erforscht.

- **Elastic ChatNoir:** Forscher der Bauhaus-Universität entwickelten *ChatNoir*, eine offene Suchoberfläche für die ClueWeb- und CC-Korpora (Source: commoncrawl.org). Dies richtet sich an die digitalen Geisteswissenschaften und die Forschung im Bereich Informationsrückgewinnung. ChatNoir indiziert Common Crawl-Seiten und bietet eine einfache Suchoberfläche, die es Nutzern ermöglicht, das CC-Archiv wie eine Suchmaschine abzufragen. Dies zeigt, dass man CC prinzipiell als „Backend“ für die Suche nutzen kann.

- **CC Search (Beta):** Common Crawl selbst startete **CC Search** (jetzt vom Creative Commons/WordPress-Team betrieben), das Nutzern die Stichwortsuche in CC ermöglicht. Die CC-Website verzeichnet Updates wie *„Große Änderungen für CC Search Beta“* Ende 2024 (verfasst von Paola Villarrela). Ziel ist es, CC-Daten zugänglicher zu machen (z.B. durch Hinzufügen von Suche nach Lizenz, Sprache usw.).

- **Startup-Vorschläge:** Die Idee einer „gemeinnützigen Suchmaschine“ wurde immer wieder diskutiert (sogar auf Hacker News (Source: dallascard.github.io). Sogar die Überschrift des Nonprofit Quarterly-Artikels lautete „Meet Common Crawl, the Nonprofit That Could Reshape the Web“ (Source: nonprofitquarterly.org). Vorerst bleibt Common Crawl selbst datenorientiert

(kein Benutzer-Suchportal), aber Dritte können darauf aufbauen. Die Existenz von CC bedeutet, dass jede gut ausgestattete Gruppe eine Suchmaschine starten könnte, *ohne das Web selbst zu crawlen*.

- **Praktische Überlegungen:** Es ist wichtig zu beachten, dass die Daten von Common Crawl Einschränkungen für die Suche aufweisen: Sie enthalten weder Page Rank, Benutzerklickdaten noch eine aktuellere Aktualität als die monatliche Granularität. Einige Websites schließen CC aus, und der Datensatz ist monatlich „eingefroren". Eine CC-basierte Engine wäre somit teilweise veraltet. Dennoch haben kleine „domänenspezifische" Suchprojekte CC erfolgreich genutzt. Zum Beispiel könnte ein Forschungsteam CC auf Nachrichten-Domains beschränken und eine spezialisierte Nachrichtensuche aufbauen. Im E-Commerce oder SEO scrapen einige Firmen CC, um offene Informationen zu Produktdaten oder Site-Rankings zu sammeln. Es wird berichtet, dass ein Blogger (Claus Matzinger von Crate.IO) über den Import von CC-Daten in eine suchfreundliche Datenbank schrieb (Source: commoncrawl.org). Wie ein langjähriger CC-Beobachter es ausdrückte: *„Wenn man nicht Google oder OpenAI oder Microsoft ist... verlässt sich fast jeder auf Common Crawl"* (Source: www.96layers.ai) für zumindest einige große Datenmengen.

Diese Fälle zeigen, dass Common Crawl **neue Arten von Diensten** ermöglicht hat, die zuvor nur Suchgiganten in Betracht ziehen konnten. Obwohl keine große kommerzielle Suchmaschine (mit Live-Abfragen) CC vollständig übernommen hat, hat das Projekt die Hürde effektiv gesenkt: Der Aufbau eines experimentellen oder akademischen Suchsystems auf Common Crawl ist unkompliziert und kostengünstig.

Datenanalyse und Forschungsergebnisse

Über Nutzungsanekdoten hinaus haben Forscher Common Crawl selbst quantitativ analysiert. Einige repräsentative Ergebnisse:

- **Datenumfang:** Ein Interview mit dem Mozilla-Forscher Stefan Baack aus dem Jahr 2024 fasste das monatliche und historische Volumen von Common Crawl zusammen (Source: www.96layers.ai). Zum Beispiel bemerkt er, dass jedes monatliche Archiv 90 TB komprimiert ist und Common Crawl in 17 Jahren *„mehr als 250 Milliarden Webseiten"* angesammelt hat (Source: www.96layers.ai). Diese Zahlen stimmen mit der offiziellen Angabe der Website von *„über 300 Milliarden Seiten"* überein (Source: commoncrawl.org). Eine solche Analyse unterstreicht die unübertroffene Größe von CC.

- **Zitationsmetriken:** Durch das Crawlen von Google Scholar oder bibliografischen Datenbanken stellte das Common Crawl-Team fest, dass ihre Daten über 10.000 Zitationen in der akademischen Literatur aufwiesen (Source: commoncrawl.org) (Source: dallascard.github.io). Dies zeigt die breite Akzeptanz in verschiedenen Bereichen. Forscher haben darauf hingewiesen, dass CC in so unterschiedlichen Bereichen wie Web-Spam-Erkennung, digitalen Bibliotheken, Journalismus (Verfolgung von Fake News) und sogar der Gesundheitsinformatik (z.B. Scannen nach medizinischer Fehlinformation) eingesetzt wird.

- **Sprach- und Site-Abdeckung:** Der Mozilla-Bericht hebt hervor, dass **Englisch** in Common Crawl **dominiert**. Er zeigt die Anzahl der Webseiten nach Land/Sprache und weist darauf hin, dass viele chinesische, japanische und Social-Media-Seiten (z.B. Facebook, Twitter) aufgrund von Crawl-Einschränkungen fehlen oder unterrepräsentiert sind (Source: www.mozillafoundation.org). Tatsächlich fehlen Seiten von Websites, die Crawler explizit blockieren. Der Bericht weist auch darauf hin, dass das Ziel von CC, die „Hassrede-Forschung" zu unterstützen, bedeutet, dass es solche Inhalte absichtlich einschließt (Source: www.mozillafoundation.org), was eine Designentscheidung ist (ungefiltert gelassen, um Analysen zu ermöglichen). Diejenigen, die an LLM-Training interessiert sind, *filtern* diese Seiten jedoch oft *heraus*.

- **Technische Robustheit:** Die Analyse von CC-Logdaten wurde durchgeführt, um das Web-Crawling selbst zu bewerten. Zum Beispiel untersuchte das Springer-Paper „Web Crawl Refusals: Insights from Common Crawl", wie Webserver Crawler blockieren oder drosseln, und nutzte dabei die eigenen Abrufprotokolle von CC (Source: commoncrawl.org). Die Ergebnisse flossen in Best Practices für das Crawling ein (z.B. wie man mit gefälschten „fake chatgpt-bot"-Blockaden umgeht).

- **Semantische Datenfülle:** Einige Projekte haben versucht, CC in großem Umfang zu annotieren. Zum Beispiel die Erstellung von Wissensgraphen durch Extrahieren von Entitäten und Beziehungen aus CC-Texten. Das CSRankings-Projekt von Stanford nutzt CC, um den Umfang der CVPR-, ICML- und NeurIPS-CS-Publikationen zu messen (obwohl das ein Nebenaspekt ist). Aber relevanter: Forscher haben CC genutzt, um offene „Common Sense"-Wissensgraphen durch das Parsen von Milliarden von Sätzen zu erstellen.

Zusammenfassend bestätigt die **Meta-Analyse** von Common Crawl dessen Umfang und Einfluss. Unabhängige Studien haben die Rohdaten der Website validiert und deren Verzerrungen untersucht. Solche Studien fließen in die Verbesserung des Datensatzes (z.B. Hervorhebung von unter-gecrawlten Bereichen des Webs) und in die Anleitung der Nutzer zur angemessenen Verwendung

(z.B. Warnungen vor Urheberrechtsproblemen) ein (Source: www.mozillafoundation.org).

Herausforderungen, Einschränkungen und Probleme

Obwohl die Daten von Common Crawl mächtig sind, sind sie nicht ohne Herausforderungen oder Kritikpunkte:

- **Bias und Repräsentativität:** Wie bereits erwähnt, ist CC sprachlich (hauptsächlich Englisch) und regional (mehr USA/EU) verzerrt. Einige Bereiche (wie afrikanische und asiatische Inhalte) sind unterrepräsentiert. Dies kann jede Analyse oder KI, die auf CC trainiert wird, verzerren. Der Mozilla-Bericht warnt ausdrücklich davor, CC *nicht* als „Ersatz für das gesamte Web“ zu behandeln (Source: www.mozillafoundation.org). Forscher ergänzen CC oft mit anderen Korpora für eine bessere Abdeckung (z.B. Nachrichten, Regierungsarchive, sprachspezifische Sammlungen).

- **Inhaltsqualität:** Common Crawl umfasst bewusst eine breite Vielfalt an Inhalten, was bedeutet, dass es auch *minderwertige, spamartige oder toxische Webseiten* erfasst. Es gibt standardmäßig keine strikte Filterung von „guten“ vs. „schlechten“ Inhalten. Für einige Anwendungsfälle (linguistische Forschung, Bias-Erkennung) ist diese Inklusivität ein Merkmal. Für das KI-Training erfordert dies jedoch eine zusätzliche Bereinigung. Zum Beispiel enthält das intelligente Paper von Ablestacks über The Pile und ähnliche Datensätze mehrere Filter, um Obszönitäten, Inhalte für Erwachsene, nicht-englischen Text usw. zu entfernen. Die Analyse von Mozilla betont, dass KI-Entwickler unerwünschte Inhalte aus CC „aussondern“ müssen, wenn ihr Ziel ein sicheres Modelltraining ist (Source: www.mozillafoundation.org). In der Praxis verwenden viele KI-Pipelines (Aleph, Redwood usw.) crowdsourced oder heuristische Listen, um CC zu filtern.

- **Urheberrecht und Lizenzierung:** Die „Nutzungsbedingungen“ von CC besagen, dass die Webseiten ohne Rücksicht auf das Urheberrecht gesammelt werden, unter der Annahme, dass Text im öffentlichen Web verwendet werden kann (ähnlich dem Betrieb von Googlebot). Der Aufstieg der KI hat jedoch rechtliche Fragen aufgeworfen. Die erwähnte Klage der New York Times deutet darauf hin, dass CC möglicherweise Tausende von urheberrechtlich geschützten Artikeln auf Nachrichtenseiten gescrapt hat, die dann in den Parametern von GPT-3 landeten. Dies verdeutlicht eine Spannung: Common Crawl ist der Ansicht, dass seine Datenerfassung rechtlich geschützt ist (z.B. unter den Ausnahmen des DMCA für Caching/Crawling oder unter der Idee der „transformative use“ im KI-Training). Rechteinhaber sind jedoch anderer Meinung. Common Crawl hat nicht explizit die Erlaubnis jedes Inhaltserstellers im Web eingeholt; es stützt sich grundsätzlich auf die Nutzungsbedingungen des Internets und robots.txt. Ende 2023 stellte Common Crawl klar, dass Inhalte, sobald sie auf CC sind, „für alle zur Nutzung bereitstehen (was Fine-Tuning und Inferenz / Retrieval-Augmentation einschließt)“ (Source: www.mozillafoundation.org). Diese Haltung ist umstritten.

- **Komitee und Governance:** Da CC ehrenamtlich betrieben wird, hängt seine Zukunft von anhaltendem Wohlwollen und der Unterstützung von Sponsoren ab. Es gibt keine garantierte Finanzierung oder große Stiftung. Wenn große Tech-Spender ihre Unterstützung zurückziehen würden, könnten die Operationen von CC gefährdet sein. Doch ab 2025 scheint das Interesse an der Erhaltung offener Webdatenprojekte hoch zu sein, angesichts des legislativen Interesses an KI-Regulierung und Open Science. Common Crawl plant (gemäß den neuesten Erklärungen), die Finanzierung zu diversifizieren und möglicherweise neue Funktionen (wie Lizenzmetadaten, Opt-out-APIs usw.) hinzuzufügen, um den Bedenken der Inhalteigentümer Rechnung zu tragen.

- **Technische Einschränkungen:** Der Datensatz ist riesig, kann aber dennoch dynamisch generierte oder hinter Formularen verborgene Inhalte übersehen. Websites, die starkes clientseitiges Rendering verwenden oder JavaScript erfordern, können für textbasierte Crawler teilweise unsichtbar sein. Einige moderne Seiten (z.B. Single-Page-Anwendungen) mit wenig statischem HTML werden möglicherweise nicht gut erfasst. Common Crawl hat mit Headless-Browsern experimentiert, aber das ist kostspielig. Daher kann CC sehr moderne, JS-lastige Websites unterindizieren. Da es außerdem nur einmal im Monat einen Durchlauf macht, kann es schnelle Updates oder kurzlebige Seiten verpassen. Nutzer, die Echtzeit-Frischdaten benötigen, können sich nicht ausschließlich auf CC verlassen.

Insgesamt erkennt das Common Crawl-Team diese Herausforderungen an. Ihre Strategie ist Transparenz: Sie veröffentlichen häufig Blogbeiträge und Antworten, um den Umfang und die Grenzen des Datensatzes zu erläutern (z.B. „Web Archiving File Formats Explained“ (Source: commoncrawl.org). Sie ermutigen die Nutzer, CC als eine **gemeinsame Infrastruktur** zu betrachten, ähnlich einem offenen Experiment, und nicht als ein perfektes Produkt.

Zukünftige Richtungen und Implikationen

Mit Blick in die Zukunft steht Common Crawl an der Schnittstelle mehrerer Trends in der Datenwissenschaft und Internet-Governance:

- **Skalierung der Datenqualität:** Common Crawl könnte fortschrittlichere Filter- oder Kennzeichnungsverfahren einführen, um den Nutzern besser zu dienen. Zum Beispiel könnte die Erstellung einer „bereinigten" Untermenge des Crawls (Entfernung von wahrscheinlichem Spam oder Inhalten für Erwachsene) die allgemeine Akzeptanz fördern. Umgekehrt könnten spezialisierte Sub-Crawls (z.B. ein mehrsprachiger Crawl oder ein qualitativ hochwertiger englischer Crawl) neue Zielgruppen anziehen.

- **Inhalteigentümer und Berechtigungen:** Während sich die Debatten um Datenrechte entwickeln, könnte Common Crawl Opt-out-Mechanismen implementieren. Bereits jetzt bieten einige Websites `DDD/Robots.txt`-Regeln für den KI-Ausschluss an. Common Crawl hat sich freiwillig bereit erklärt, `x-robot-tags` zu respektieren, die jegliches Nicht-Bot-Crawling blockieren (im DRM-Stil). Zukünftige Systeme könnten es Website-Betreibern ermöglichen, die Entfernung aus dem CC-Archiv zu beantragen. Andererseits gefährden solche Opt-outs die Einheitlichkeit von Datensätzen für Forscher. Das Projekt wird wahrscheinlich weiterhin mit Rechtsexperten zusammenarbeiten, um ein Gleichgewicht zu finden.

- **Initiativen für offene Suche:** Es gibt eine wachsende Befürwortung von „Suchinfrastruktur als öffentliches Gut". Common Crawl könnte die Datengrundlage einer neuen Generation offener Suchmaschinen oder Wissensgraphen werden. Zum Beispiel spiegeln Projekte wie **OpenWebIndex** (ein vorgeschlagenes EU-finanziertes Projekt) die Mission von Common Crawl wider. Wir könnten Partnerschaften sehen, bei denen der Crawl von Common Crawl spezialisierte Indizes antreibt (z.B. eine akademische Suchmaschine für Bildungsinhalte oder eine offene Einkaufssuche). Die Veröffentlichung der **Index API** von Common Crawl (angekündigt 2023) zeigt eine Bewegung in diese Richtung.

- **KI und verantwortungsvolle Nutzung:** Da die Daten von Common Crawl generative KI antreiben, könnte die Stiftung in „KI-Ethik"-Funktionen investieren. Dies könnte Anmerkungen (Kennzeichnung von Seiten, die Propaganda oder Gesundheitsdesinformationen enthalten) oder die Integration von De-Biasing-Filtern umfassen. Der Mozilla-Bericht schlägt vor, dass Entwickler „robuste Datenfilter" hinzufügen sollten (Source: www.mozillafoundation.org); Common Crawl selbst könnte beginnen, vorgefilterte Versionen oder Tools zum Filtern anzubieten (z.B. einen Toxizitätsfilter).

- **Weitere Analysen durch Common Crawl:** Die Stiftung könnte mehr Datenanalysen intern erstellen. Zum Beispiel zeigen ihre GitHub-Dashboards „Crawl Stats" und „Graph Stats". Eine Erweiterung dieser Dashboards, um Echtzeit-Sprachaufschlüsselungen, Metriken zur Domain-Vielfalt oder semantische Trends anzuzeigen, könnte wertvoll sein. Dies würde sowohl Nutzern als auch Geldgebern helfen, den Umfang der Ressource zu verstehen.

- **Globale Partnerschaften:** Um die Abdeckung zu verbessern, könnte Common Crawl mit internationalen Universitäten oder NGOs zusammenarbeiten, um den Crawl mit mehr globalen Inhalten zu bestücken (z.B. durch länderspezifische Top-100-Domains). Es könnte auch mit nationalen Bibliotheken (wie Europeana oder nationalen Webarchiven) zusammenarbeiten, um „Walled Gardens" des Webs zu integrieren.

Im weiteren Sinne deutet die Wirkung von Common Crawl darauf hin, dass **Daten-Commons** (offene Dateninfrastruktur) ein tragfähiges Modell für andere Bereiche sein könnten: man stelle sich offene Korpora wissenschaftlicher Arbeiten, Bilder oder Umweltsensoren vor. Der Erfolg von Common Crawl bietet eine Vorlage: minimales Team, Cloud-Sponsoren, offene Daten. Es zeigt, dass unter den richtigen Bedingungen *„Daten die neue öffentliche Infrastruktur sind."*

Fazit

Common Crawl entstand aus Gil Elbaz' Vision eines offenen Web-Index und hat sich in fast zwei Jahrzehnten zu einer zentralen Ressource für datengesteuerte Innovationen entwickelt. Seine **Geschichte** ist eine Geschichte bescheidener Anfänge (eine kleine gemeinnützige Organisation im Jahr 2007), die durch Gemeinschaftsanstrengungen und Cloud-Unterstützung zu einem **gigantischen Webarchiv** heranwuchs (Source: nonprofitquarterly.org) (Source: www.96layers.ai). Es entstand aus dem Engagement für offene Daten und hat sich an dieses Prinzip gehalten: webweite Informationen demokratisch zugänglich zu machen, nicht proprietär.

Heute wird Common Crawl von Tausenden von Forschern und Entwicklern weltweit genutzt. Es treibt die Spitze der KI an (praktisch alle großen Sprachmodelle verlassen sich darauf) und ermöglicht Start-ups, die sich sonst die Infrastruktur von Google nicht leisten könnten. Tabelle 2 in diesem Bericht zeigte einige konkrete Beispiele, aber eine umfassende Aufzählung wäre noch länger. Seine

Präsenz in über **10.000** akademischen Publikationen (Source: commoncrawl.org) (Source: dallascard.github.io) ist ein Beweis für seinen Einfluss.

Doch mit großer Macht kommen auch große Verantwortung und Komplikationen. Die Nutzung von Common Crawl im KI-Training hat soziale und rechtliche Fragen aufgeworfen – insbesondere da generative Modelle den öffentlichen Diskurs prägen. Das Common Crawl-Team ist sich dessen bewusst und hat sich mit der Community darüber ausgetauscht, wie die Daten verantwortungsvoll genutzt werden können. Der Mozilla-Bericht und andere Analysen deuten darauf hin, dass CC noch Jahre lang Teil der Debatten über KI-Ethik und Urheberrecht sein wird (Source: www.mozillafoundation.org) (Source: www.mozillafoundation.org).

Mit Blick in die Zukunft scheint der Kurs von Common Crawl auf eine fortgesetzte Expansion und tiefere Integration in die offene Forschung ausgerichtet zu sein. Da die Rechenleistung wächst und KI immer mehr Daten sucht, wird der Wert des offenen Webarchivs von Common Crawl wahrscheinlich steigen. Die Gemeinschaft um es herum könnte wachsen und sich möglicherweise von einem kleinen Team zu einem größeren kollaborativen Konsortium entwickeln. Es gibt aufkommende Projekte zur Erweiterung seiner Fähigkeiten (wie reichhaltigere Suchindizes oder Filteroptionen), die die Ära von „Search 2.0“ prägen könnten (Source: commoncrawl.org).

Zusammenfassend ist die **vollständige Geschichte** von Common Crawl eine Fallstudie, wie eine kleine, zielgerichtete Initiative die **Daten-Commons** dramatisch öffnen kann. Es begann als Reaktion auf Monopolängste in der Websuche und hat tatsächlich Türen für Innovationen geöffnet. Sein Gründer Gil Elbaz und seine Mitarbeiter schufen erfolgreich „das Web als riesige Datenbank“, zugänglich für alle (Source: nonprofitquarterly.org). Die Geschichte von Common Crawl – vom ersten Fünf-Milliarden-Seiten-Crawl bis zu Tausenden von Milliarden Seiten heute – veranschaulicht die Kraft offener Infrastruktur. Seine zukünftige Rolle wird sich wahrscheinlich vertiefen, während die Gesellschaft sich mit den Vorteilen und Herausforderungen von webbasierter KI und Open Science auseinandersetzt.

Alle oben genannten Behauptungen werden durch zitierte Quellen aus der eigenen Dokumentation von Common Crawl, Medienberichten, Interviews und wissenschaftlichen Analysen gestützt (Source: commoncrawl.org) (Source: www.96layers.ai) (Source: www.mozillafoundation.org) (Source: dallascard.github.io). Diese Referenzen bilden eine transparente Grundlage für die Behauptungen des Berichts. Indem wir mehrere Perspektiven (technisch, organisatorisch, ethisch) abdecken und quantitative Daten (Seitenanzahl, Nutzungsstatistiken, zitierte Veröffentlichungen) einbeziehen, haben wir uns um eine **gründliche, evidenzbasierte Darstellung** der Geschichte, des aktuellen Status und der zukünftigen Implikationen von Common Crawl bemüht.

Tags: common-crawl, web-crawling, llm-trainingsdaten, offene-daten, gil-elbaz, big-data, web-repository, apache-nutch