¿Qué es Common Crawl? Una historia del conjunto de datos de la web abierta

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

Resumen Ejecutivo

Common Crawl es una fundación sin fines de lucro **501(c)(3)** (fundada en 2007) que mantiene un **repositorio gratuito y abierto de datos de rastreo web** (Source: commoncrawl.org) (Source: commoncrawl.org). Su misión es *democratizar el acceso a la información web* proporcionando **conjuntos de datos de rastreo web a escala de petabytes** sin costo alguno. Durante los últimos más de 15 años, Common Crawl ha recopilado del orden de **300-400 mil millones de páginas web**, abarcando más de 15 años de rastreo continuo (Source: commoncrawl.org) (Source: www.96layers.ai). Cada mes añade aproximadamente **3-5 mil millones de páginas nuevas** (alrededor de 90 TB comprimidos, ~400 TB sin comprimir) (Source: www.96layers.ai) (Source: commoncrawl.org). Aunque comenzó como un proyecto minúsculo (solo unos pocos empleados) (Source: www.96layers.ai), el corpus disponible públicamente de Common Crawl ahora sustenta una amplia gama de usos comerciales y de investigación. En particular, suministra la mayor parte de los datos de entrenamiento para los modernos modelos de lenguaje grandes (LLM) – por ejemplo, **más del 80% de los tokens en GPT-3 de OpenAI** provienen de datos de Common Crawl (Source: www.mozillafoundation.org) – y es citado en más de **10,000 publicaciones académicas** (Source: commoncrawl.org) (Source: dallascard.github.io). Ha permitido startups (por ejemplo, TinEye, Lucky Oyster) y proyectos de investigación (por ejemplo, incrustaciones de palabras GloVe, análisis de censura web) que de otro modo carecerían de los recursos para rastrear toda la web. Common Crawl sirve así como una "infraestructura neutral y sin fines de lucro" para datos web (Source: www.96layers.ai), igualando las condiciones para que incluso pequeñas organizaciones e investigadores puedan acceder a información a escala web.

Este informe proporciona una **historia y análisis exhaustivos de Common Crawl**. Cubre los orígenes del proyecto (motivaciones clave, antecedentes del fundador, desarrollo temprano), estructura organizativa y financiación, métodos y tecnología de recopilación de datos, crecimiento del conjunto de datos y las **múltiples formas en que se utilizan los datos hoy en día** (en entrenamiento de IA/LLM, investigación académica, productos industriales, etc.). Examinaremos el contexto social y técnico (por ejemplo, el dominio de Google y la necesidad de datos web abiertos), resumiremos **estadísticas cuantitativas** (páginas

recopiladas, volumen de datos, recuentos de citas) y presentaremos estudios de caso que ilustran el impacto de Common Crawl. También discutimos desafíos (sesgo de cobertura, problemas de derechos de autor) y direcciones futuras. Todas las afirmaciones y hechos están respaldados por fuentes autorizadas de la organización Common Crawl, medios de comunicación, entrevistas y publicaciones de investigación.

Introducción y Antecedentes

La **World Wide Web** ha crecido hasta convertirse en un vasto ecosistema de información descentralizado. Los motores de búsqueda modernos como Google y Bing rastrean continuamente la web para crear sus propios índices, pero estos índices son propietarios. A mediados de la década de 2000, **no existía ningún repositorio importante de datos de rastreo web de acceso público** para personas ajenas. Solo unas pocas organizaciones —notablemente la organización sin fines de lucro Internet Archive— intentaron preservar páginas web (por ejemplo, a través de la Wayback Machine). Sin embargo, la *Wayback Machine* de Internet Archive está diseñada para el archivo y la navegación de instantáneas de páginas web a lo largo del tiempo bajo demanda; no está optimizada para el análisis de datos a gran escala o la minería algorítmica del contenido de la web (Source: dallascard.github.io).

En este contexto, la idea de construir un **"índice web abierto"** comenzó a surgir. Empresarios e investigadores reconocieron que solo las empresas más grandes (Google, Microsoft, Yahoo, Baidu, etc.) tenían los recursos para rastrear miles de millones de páginas con alta frecuencia, dejando a los actores más pequeños sin acceso a estos datos brutos. Por ejemplo, los investigadores universitarios y las startups a menudo necesitaban grandes corpus web para tareas de procesamiento de lenguaje natural (NLP), minería de datos y aprendizaje automático, pero carecían de los medios para rastrear toda la web por sí mismos. Un repositorio abierto de datos de rastreo web **democratizaría el acceso** y fomentaría la innovación, de manera similar a cómo los conjuntos de datos abiertos (por ejemplo, Wikipedia) impulsaron nuevas investigaciones.

Common Crawl fue concebido y lanzado para satisfacer esta necesidad. Su fundador, **Gil Elbaz**, es un emprendedor en serie y tecnólogo: a finales de los años 90 cofundó Applied Semantics (la empresa que construyó la tecnología más tarde conocida como Google AdSense) (Source: www.96layers.ai) (Source: www.96layers.ai). Después de que Google adquiriera Applied Semantics, Elbaz trabajó en Google hasta 2007. En entrevistas, explicó que su partida fue motivada por la preocupación sobre la concentración de datos y su impacto en la innovación. Consideraba el rastreo propietario masivo de Google como clave para su monopolio en la innovación de búsqueda (Source: www.96layers.ai) (Source: www.96layers.ai). Para contrarrestar esto, Elbaz imaginó "empresas de datos neutrales" —proyectos de infraestructura abiertos y sin fines de lucro que rastrearían la web y proporcionarían los datos **gratuitamente** a investigadores y empresas. Uno de esos proyectos fue **Common Crawl**, fundado en 2007. Como dijo Elbaz:

> *"Common Crawl estaba destinado a ser como una infraestructura neutral sin fines de lucro que debería imitar la forma en que Google rastreaba la web... y luego poner esos datos a disposición de cualquiera de forma gratuita, para nivelar el campo de juego del desarrollo tecnológico" (Source: www.96layers.ai).*

La motivación de Elbaz, por lo tanto, fue explícitamente **nivelar el campo de juego**. Quería que las pequeñas startups y los investigadores académicos tuvieran la misma información de "índice de búsqueda" en bruto que tenía Google, para que la innovación no fuera monopolizada por una sola empresa (Source: www.novaspivack.com) (Source: www.96layers.ai). Esta visión resonó con otros líderes de la comunidad web abierta. Tecnólogos prominentes como **Nova Spivack** (un emprendedor temprano de Internet) y **Carl Malamud** (un pionero de los datos gubernamentales abiertos) se unieron a la junta directiva fundadora de Common Crawl (Source: www.novaspivack.com). Con el tiempo, la junta asesora creció para incluir luminarias como el director de investigación de Google **Peter Norvig** y el director del MIT Media Lab **Joi Ito** (Source: www.thekurzweillibrary.com) (Source: commoncrawl.org), lo que subraya la prominencia del proyecto.

En pocos años, Common Crawl se había convertido en una fundación independiente sin fines de lucro. Desde su lanzamiento, se registró como una organización 501(c)(3) de California, la **Common Crawl Foundation** (Source: commoncrawl.org) (Source: commoncrawl.org). Su declaración de misión es concisa: *"democratizar el acceso a la información web produciendo y manteniendo un rastreo abierto de la web"*. La página de inicio de Common Crawl lo describe como *"un repositorio gratuito y abierto de datos de rastreo web que puede ser utilizado por cualquiera"* (Source: commoncrawl.org). Gil Elbaz se desempeñó como Presidente de la Junta y a menudo se le atribuye ser el fundador del proyecto (Source: commoncrawl.org) (Source: www.novaspivack.com). Otros miembros clave del equipo inicial incluyeron al ingeniero principal **Ahad Rana** y más tarde a la directora **Lisa Green** (anteriormente de Creative Commons) (Source: www.novaspivack.com).

Estructura Organizativa y Financiación

Common Crawl opera como una pequeña organización sin fines de lucro. Su página de inicio y páginas de equipo de 2025 indican que el equipo central ha sido históricamente muy pequeño, literalmente "menos de cinco personas" en los primeros años (Source: www.96layers.ai). Por ejemplo, a principios de la década de 2010, el proyecto funcionaba con solo un puñado de ingenieros y voluntarios. Incluso en el momento en que OpenAI publicó el artículo de GPT-3 en 2020, Common Crawl supuestamente tenía solo **un empleado a tiempo completo** (Source: www.96layers.ai) (aunque para 2025 el equipo es más grande). Gil Elbaz funciona como Presidente (y fue copresidente de Factual/Foursquare), y nombres como Peter Norvig son asesores (Source: commoncrawl.org). Sin embargo, las operaciones diarias dependen de un pequeño personal permanente y de las contribuciones de voluntarios y colaboradores.

La organización se financia principalmente a través de **donaciones y patrocinios**, especialmente de proveedores de la nube. Desde 2012 en adelante, Amazon Web Services (AWS) ha alojado los datos de Common Crawl sin costo alguno bajo el programa AWS Public Datasets (Source: alchetron.com). El patrocinio de datos públicos de AWS proporciona el inmenso almacenamiento requerido (muchos cientos de terabytes) sin cobrar a Common Crawl. Otras plataformas en la nube (por ejemplo, Microsoft Azure, Google Cloud) también pueden estar involucradas en los archivos, pero AWS es el anfitrión principal. Además, empresas como Amazon han ofrecido concursos de pequeñas subvenciones (por ejemplo, créditos de AWS de $50) para fomentar el uso de los datos (Source: commoncrawl.org). Es probable que la fundación también reciba modestas donaciones filantrópicas, aunque **Common Crawl nunca ha aceptado inversión de capital de riesgo ni ha funcionado como una empresa comercial**. (Deliberadamente sigue siendo una organización sin fines de lucro para mantenerse "neutral" y libre de motivos de lucro (Source: www.novaspivack.com) (Source: www.96layers.ai).)

En resumen, Common Crawl es el producto colaborativo de unos pocos tecnólogos apasionados y el ecosistema de la computación en la nube. Sus costos operativos relativamente bajos (porque evita las tarifas de almacenamiento) le permiten persistir con una financiación mínima. A partir de 2024, Common Crawl sigue siendo "en gran parte desconocido para el público en general", pero se le reconoce por desempeñar "un papel importante" en campos como la IA generativa (Source: www.mozillafoundation.org). El informe de la Fundación Mozilla de 2024 enfatiza que Common Crawl es "una pequeña organización sin fines de lucro" con un impacto masivo (Source: www.mozillafoundation.org).

Recopilación de Datos: Rastreo y Tecnología

Common Crawl ejecuta un rastreador web automatizado (llamado **CCBot**) que escanea continuamente la web pública para construir su conjunto de datos. El rastreador está construido sobre el framework de código abierto Apache Nutch, que maneja el descubrimiento de URL, la obtención de páginas y el seguimiento de hipervínculos (Source: datadome.co). (De hecho, en 2013 Common Crawl cambió a usar Apache Nutch como su rastreador principal "en lugar de un rastreador personalizado" (Source: alchetron.com), y migró del formato de archivo "ARC" más antiguo al formato estándar **WARC** al mismo tiempo (Source: alchetron.com).) CCBot se identifica en el user-agent como “CCBot/2.0” (Source: datadome.co), aunque se desaconseja confiar únicamente en la cadena del user-agent porque los bots pueden falsificar identidades. CCBot rastrea desde direcciones IP de Amazon AWS. En años anteriores, los rangos de IP de CCBot estaban documentados públicamente (por ejemplo, `38.107.191.66` – `38.107.191.119`) (Source: datadome.co), pero ahora el rastreador está completamente basado en la nube.

Robots.txt y ética: Como los rastreadores de buena ciudadanía, CCBot **respeta las reglas de `robots.txt` y las etiquetas nofollow** (Source: alchetron.com), por lo que evita las páginas explícitamente no permitidas por los propietarios del sitio. Se concentra en contenido de acceso público (páginas HTML) y almacena el contenido de la página en bruto (HTML y texto) en los archivos de rastreo. A diferencia de Internet Archive, que busca preservar páginas con el fin de archivar y reproducir (incluyendo imágenes, scripts y comportamientos del lado del cliente) (Source: dallascard.github.io), el enfoque de Common Crawl está en el contenido textual y los metadatos útiles para la minería de datos y el aprendizaje automático. Específicamente, Common Crawl *no* almacena ni analiza imágenes, videos, CSS u otros recursos estáticos en detalle; el énfasis está en el *texto HTML en bruto y los metadatos asociados*. Esto hace que el corpus de Common Crawl sea más directamente útil para NLP y análisis de datos, a expensas de una instantánea visual completa.

Metodología de rastreo: Common Crawl suele realizar un **rastreo de un mes de duración**, lo que significa que ejecuta CCBot continuamente para obtener páginas durante aproximadamente un mes, luego publica los resultados como un "archivo de rastreo". Repite esto aproximadamente cada mes. Históricamente, el cronograma ha variado: en los primeros años hubo alrededor de 4 rastreos por año (Source: alchetron.com), pero luego se volvió mensual. Cada rastreo mensual comienza a partir de un enorme

conjunto de URL semilla (puntos de entrada iniciales) en la web pública y sigue enlaces para descubrir nuevas URL, podando en el camino utilizando heurísticas basadas en dominios para mantener una amplia cobertura. El resultado de cada rastreo es una colección de archivos WARC (archivos comprimidos de páginas obtenidas) más metadatos adjuntos (por ejemplo, tablas de URL, extractos de texto, gráficos de enlaces) (Source: alchetron.com). Alrededor de mediados de 2012, Common Crawl también comenzó a publicar texto y metadatos extraídos de cada rastreo, en lugar de solo WARC en bruto (Source: alchetron.com).

Escala y crecimiento: La escala de la operación de Common Crawl es masiva. Según una entrevista de 2023, *cada mes* Common Crawl recopila **entre 3 y 5 mil millones de páginas web**, lo que equivale a "500 veces más páginas web que [toda Wikipedia]" (Source: www.96layers.ai). Los datos mensuales comprimidos son del orden de **90 terabytes** (aproximadamente 400 TB sin comprimir) (Source: www.96layers.ai). A lo largo de más de una década, Common Crawl ha acumulado **cientos de miles de millones de páginas**. En un informe (abril de 2024), se señaló que "a lo largo de sus 17 años de historia, Common Crawl ha recopilado más de *250 mil millones de páginas web*" (Source: www.96layers.ai). Su propia página de inicio (a finales de 2025) afirma "*más de 300 mil millones de páginas a lo largo de 15 años*" (Source: commoncrawl.org). (Estas cifras son ampliamente consistentes, dado el rastreo continuo). Para contextualizar, en su lanzamiento a principios de 2013, el conjunto de datos inaugural de Common Crawl comprendía aproximadamente **5 mil millones de páginas** (≈81 terabytes) (Source: nonprofitquarterly.org) (Source: www.thekurzweillibrary.com). A mediados de 2015, los rastreos archivados cubrían aproximadamente **1.8 mil millones de páginas (145 TB)** a lo largo de 4 rastreos anuales (Source: alchetron.com). Hoy en día, solo el rastreo mensual supera esos totales anteriores.

Además del contenido de las páginas, Common Crawl también publica **grafos de enlaces a nivel de host y dominio** y otros conjuntos de datos derivados (por ejemplo, URLs que contienen una consulta determinada, o aproximaciones de PageRank a nivel de dominio). Estos están disponibles en su página *Data* y en GitHub, y se actualizan regularmente. Los archivos WARC sin procesar y el texto procesado se alojan en **Amazon S3** (Conjunto de Datos Público de AWS) y sitios espejo. Los usuarios pueden descargar rastreos específicos por mes/año mediante HTTP o utilizar herramientas de big data (por ejemplo, Amazon Athena, Spark) para consultar los datos en el lugar. Common Crawl también proporciona herramientas de ayuda e índices (por ejemplo, un índice de URL) para facilitar la búsqueda de páginas de interés.

En general, la tecnología de rastreo de Common Crawl ha evolucionado pero ha permanecido abierta. Utiliza componentes estándar y conocidos (Apache Nutch, la nube de Amazon) y código de código abierto para el procesamiento de datos. Debido a que es un proyecto sin fines de lucro, aprovecha la nube de formas creativas: evita pagar costos de almacenamiento al permanecer en el nivel gratuito de AWS, y elude las tarifas de transmisión de datos (egress) al fomentar el análisis en la plataforma de AWS. La infraestructura central de Common Crawl es relativamente simple, pero el resultado es enorme: terabytes de datos web abiertos agregados y mantenidos como un recurso común (Source: www.96layers.ai) (Source: dallascard.github.io).

Conjunto de Datos y Estadísticas

El conjunto de datos público de Common Crawl es uno de los corpus de texto más grandes existentes, comparable en escala al almacenamiento de los principales motores de búsqueda. Las estadísticas clave sobre el corpus (a mediados de 2025) son:

- **Tamaño del corpus:** Más de *300 mil millones de páginas web únicas* (documentos HTML) recopiladas (Source: commoncrawl.org). (En comparación, esto es miles de veces más grande que toda la Wikipedia en inglés).
- **Lapso temporal:** Instantáneas mensuales desde 2008 o 2009 hasta el presente (más de 15 años) (Source: commoncrawl.org). Cada instantánea suele contener páginas rastreadas en ese mes. La colección crece de forma aditiva cada año.
- **Tasa de crecimiento mensual:** Típicamente **3-5 mil millones de páginas por mes**, lo que produce aproximadamente **90 TB comprimidos (~400 TB sin comprimir)** cada mes (Source: www.96layers.ai) (Source: commoncrawl.org). En un año, eso es del orden de **30-60 mil millones de páginas** y cientos de terabytes.
- **Frecuencia de rastreo:** Generalmente un rastreo por mes (aunque al principio era menos). El archivo es acumulativo en el sentido de que cada rastreo es una nueva instantánea, pero en la práctica los usuarios pueden combinar datos de varios meses.
- **Volumen de datos:** Cientos de terabytes por rastreo distribuidos en archivos WARC, más texto derivado y metadatos en archivos adyacentes. Por ejemplo, el rastreo inaugural de 2013 fue de 81 TB (Source: nonprofitquarterly.org), y los rastreos modernos son más grandes. En total, los archivos de Common Crawl ascienden a **múltiples petabytes** de datos comprimidos (el informe de Mozilla de 2024 cita "más de 9.5 petabytes" de datos de Common Crawl) (Source: www.mozillafoundation.org).

- **Uso en la literatura de investigación:** Más de **10,000 artículos de investigación** han citado a Common Crawl como fuente de datos (Source: commoncrawl.org) (Source: dallascard.github.io). Esta cifra parece haberse duplicado aproximadamente cada pocos años. (El número exacto es difícil de verificar, pero el sitio web afirma con orgullo "citado en más de 10,000 artículos de investigación" (Source: commoncrawl.org), y datos independientes muestran que el recuento era mucho menor en 2013).

Estas cifras aproximadas demuestran la escala masiva de los datos. Es notable que solo unas pocas organizaciones privadas (Google, Microsoft, Amazon, Facebook) tienen una capacidad de rastreo a escala web comparable, y mantienen los datos como propietarios. Por el contrario, el archivo de Common Crawl está listado públicamente en AWS Open Data y otros espejos, lo que permite a **cualquiera** descargarlo o analizarlo (Source: registry.opendata.aws).

Es importante destacar que Common Crawl deja claro que su conjunto de datos **no** es la "web completa" ni se garantiza que sea exhaustivo. La cobertura está sesgada hacia las páginas web accesibles en inglés (los sitios bloqueados a través de robots.txt están excluidos, y las principales plataformas como Facebook bloquean el rastreo). Un estudio de Mozilla de 2024 advirtió expresamente que *"Tratar acríticamente a Common Crawl como una 'copia de la web' declara una sección relativamente pequeña de páginas web principalmente en inglés como representativa del mundo entero."* (Source: www.mozillafoundation.org). En la práctica, Common Crawl representa la **"web visible"** (la parte accesible desde enlaces HTML típicos) a partir de cada fecha de rastreo, con énfasis en la diversidad (no se centra exclusivamente en los dominios principales) y la frescura.

A pesar de las limitaciones, la gran amplitud de los datos de Common Crawl los hace extremadamente valiosos. **Supera con creces** cualquier conjunto de datos estático que la mayoría de los investigadores podrían recopilar por sí mismos. Los modelos de lenguaje natural modernos utilizan comúnmente **cientos de miles de millones de palabras** de Common Crawl. Por ejemplo, la incrustación de palabras GloVe de Stanford (2014) fue entrenada con **840 mil millones de tokens** extraídos de Common Crawl (Source: huggingface.co). Y los principales LLM ingieren rutinariamente miles de páginas web informales de Common Crawl (como se detalla a continuación). Los datos también se utilizan en el análisis de grafos web, la investigación de recuperación de información (por ejemplo, la construcción de motores de búsqueda para el conjunto de datos ClueWeb (Source: commoncrawl.org), y la minería específica de dominios (como la extracción de texto paralelo para la traducción automática (Source: huggingface.co).

La Tabla 1 a continuación resume algunas de estas métricas y hechos clave:

MÉTRICA/HECHO	VALOR/DESCRIPCIÓN	FUENTE
Año de fundación	2007 (establecida como una organización sin fines de lucro 501(c)(3) en 2007)	[9†L0-L4], [7†L19-L24]
Fundador y Presidente	Gil Elbaz (tecnólogo, cofundador de Applied Semantics/AdSense)	[47†L0-L4], [6†L144-L152]
Junta Asesora (notable)	Peter Norvig de Google, Joi Ito del MIT, Nova Spivack, Carl Malamud	[30†L36-L38], [47†L19-L24], [45†L10-L18]
Tipo de organización	Organización sin fines de lucro 501(c)(3) (California)	[9†L0-L4], [7†L19-L24]
Antigüedad/alcance del conjunto de datos	2008/2009 – presente (más de 15 años de páginas web mensuales)	[9†L10-L17], [2†L20-L24]
Páginas totales recopiladas	**~300+ mil millones** de páginas web (acumulativas)	[9†L10-L17], [2†L20-L24]
Crecimiento mensual (páginas)	~3-5 mil millones de páginas nuevas añadidas por mes (promedio)	[2†L20-L24], [9†L14-L17]
Tamaño de datos mensual	~90 terabytes comprimidos (~400 TB sin comprimir) por rastreo mensual	[2†L20-L24]
Criterios de inclusión	Páginas HTML públicas (obedeciendo robots.txt); enfoque en texto sin procesar (sin imágenes/videos).	[52†L22-L31], [19†L28-L31]
Usos notables del proyecto	Entrenamiento de IA/ML (GPT-3, PaLM, etc.), incrustaciones de palabras (GloVe 840B tokens), corpus de investigación (C4, The Pile), motores de búsqueda	[60†L23-L30], [61†L32-L39], [52†L49-L57]
Citas de investigación (aprox.)	**>10,000** artículos publicados que citan a Common Crawl	[9†L12-L17], [52†L34-L40]
Conjunto de datos alojado en Amazon	Alojado a través de AWS Open Data (gratuito para usuarios a través de S3/Athena/AWS)	[19†L33-L39], [25†L12-L19] (registro de AWS)
Mayor cobertura de LLM	~80-85% de los tokens de entrenamiento de GPT-3 provienen de Common Crawl (Source: www.mozillafoundation.org); ~64% de los LLM encuestados (2019-2023) usan CC (Source: www.mozillafoundation.org).	

(Tabla 1: Hechos y estadísticas clave sobre Common Crawl, con fuentes citadas.)

Historia y Desarrollo

El desarrollo de Common Crawl puede verse cronológicamente a través de varios hitos clave:

- **2007 – Inicio del Proyecto:** Gil Elbaz "se me acercó con una visión ambiciosa: quería crear un rastreo abierto y sin fines de lucro de la Web" (Source: www.novaspivack.com). En 2007 fundó oficialmente la Common Crawl Foundation. Los primeros colaboradores incluyeron a Nova Spivack y Carl Malamud, quienes se convirtieron en miembros de la junta (Source: www.novaspivack.com). En esta etapa solo había personas trabajando en ello (el propio Elbaz, Ahad Rana como ingeniero principal, algunos voluntarios). Spivack relata: *"Gil y el ingeniero principal, Ahad Rana, se pusieron a trabajar en la construcción del proyecto."* (Source: www.novaspivack.com). El objetivo era crear "el primer índice de búsqueda de la Web verdaderamente abierto, sin fines de lucro, de 5 mil millones de páginas" (Source: www.novaspivack.com). (De hecho, los primeros datos de rastreo publicados alrededor de 2013 contenían aproximadamente 5 mil millones de páginas, 81 TB, según informó MIT Tech Review (Source: nonprofitquarterly.org) (Source: www.thekurzweillibrary.com).)

- **2008–2011 – Primeros Rastros:** Tras su inicio, Common Crawl comenzó rastreos mensuales (~trimestrales) de una parte de la web. En esos años, los volúmenes de datos eran menores; las primeras publicaciones de blog indican solo unos pocos terabytes por rastreo. El énfasis estaba en construir la tubería (rastreador basado en Nutch, archivos WARC, procesos simples de Hadoop para extraer texto). Inicialmente, el equipo escribió código personalizado, pero en 2013 anunciaron el cambio a Apache Nutch y la adopción del formato de archivo WARC para todos los datos de rastreo (Source: alchetron.com). El uso de Amazon S3 para el almacenamiento probablemente comenzó en esta época.

- **2012 – Asociación con Amazon AWS:** Un punto de inflexión importante ocurrió en 2012 cuando **Amazon Web Services** aceptó a Common Crawl en su programa de Conjuntos de Datos Públicos (Source: alchetron.com). AWS acordó alojar los archivos de rastreo en su nube sin costo. Esto fue crucial, ya que permitió a Common Crawl escalar de gigabytes a petabytes sin incurrir en gastos de almacenamiento. (Paralelamente, AWS y Common Crawl colaboraron más tarde en concursos; por ejemplo, AWS ofreció a los participantes del concurso $50 en créditos para usar los datos (Source: commoncrawl.org).) También a finales de 2012, la empresa de motores de búsqueda *Blekko* donó metadatos de sus propios rastreos (febrero-octubre de 2012) a Common Crawl (Source: alchetron.com). Los registros de Blekko ayudaron a mejorar la cobertura del rastreo y a reducir las páginas no deseadas (spam, pornografía, manipulaciones de SEO) (Source: alchetron.com).

- **2013 – Lanzamiento Formal y Reconocimiento:** A principios de 2013, el primer gran lanzamiento público de Common Crawl (el "índice de 5 mil millones de páginas") atrajo la atención de los medios. MIT Technology Review (a través del blog de Ray Kurzweil) publicó una historia en enero de 2013 titulada "Una base de datos gratuita de toda la Web podría generar el próximo Google" (Source: www.thekurzweillibrary.com). La historia destacaba que *"Common Crawl ofrece más de cinco mil millones de páginas web, disponibles de forma gratuita para que investigadores y emprendedores puedan probar cosas que de otro modo solo serían posibles para aquellos con acceso a los recursos de Google."* (Source: www.thekurzweillibrary.com). Para entonces, Peter Norvig y Joi Ito se habían unido a la junta asesora (Source: www.thekurzweillibrary.com). Se lanzó el propio sitio y panel de control de Common Crawl, anunciando el archivo de datos de una década y obteniendo los primeros usuarios de investigación.

- **2014–2019 – Expansión de Datos y Crecimiento del Ecosistema:** A mediados de la década de 2010, Common Crawl continuó con los rastreos mensuales, y el conjunto de datos acumulativo creció rápidamente. Cada año, se construyó más investigación y desarrollo sobre estos datos. Los eventos importantes incluyen:
  - **2014–2015:** Extracción de datos estructurados: Common Crawl comenzó a extraer texto y metadatos de las páginas sin procesar y a publicarlos junto con los archivos WARC. Se pusieron a disposición datos para idiomas como español, alemán, etc. La comunidad también desarrolló herramientas para consultar los datos en el lugar, como Recipes e Index (a través de AWS Athena).
  - **2016:** Introducción de **CCBot v2.0** con un agente de usuario actualizado (Source: datadome.co) y mejoras en el cumplimiento de robots.txt. El papel de Common Crawl en la investigación se consolidó a medida que tareas de PNL como GloVe (84 GB) utilizaron datos de CC (Source: huggingface.co).

- **2017–2019:** El conjunto de datos superó las decenas de miles de millones de páginas. Durante este tiempo, Europa inició el *Norvig Web Data Science Award* (apoyado por Common Crawl y SURFSara), fomentando el uso académico de los datos. Además, el equipo de ingeniería central siguió siendo pequeño; en entrevistas señalaron tener tan solo 3 empleados alrededor de 2017 (Source: www.96layers.ai). Para 2019, Common Crawl fue reconocido como una fuente clave para el entrenamiento de modelos neuronales, aunque todavía pasaba desapercibido para el público en general.

- **2020–2022 – Auge de la IA:** El auge de la IA en la era COVID puso a Common Crawl en el centro de atención. GPT-3 de OpenAI (publicado a mediados de 2020) utilizó Common Crawl como fuente de datos principal. Equipos de investigación detrás de modelos como Grover (Zellers et al., 2019) se entrenaron explícitamente en CC para la generación de noticias falsas (Source: dallascard.github.io). RoBERTa (2019) de Meta y T5 de Google también se basaron en corpus derivados de CC. En 2020, los datos de Common Crawl se incorporaron a grandes conjuntos de datos de investigación como "C4" (utilizado para T5) y "The Pile" (un corpus en inglés de 800 GB), ambos reconociendo públicamente a CC como un componente importante (Source: dallascard.github.io). El público comenzó a oír hablar de "billones de tokens" extraídos de la web para la IA, y Common Crawl fue identificado como una fuente clave. Sin embargo, Common Crawl en sí mismo siguió siendo pequeño; se informó que para cuando se lanzó GPT-3, la organización posiblemente tenía solo un empleado trabajando en ello (Source: www.96layers.ai).
- **2023–2025 – Era Actual y Reconocimiento Público:** En 2023 y 2024, Common Crawl experimentó un aumento de la atención pública debido a dos factores: (a) el auge de la IA generativa, para la cual los datos abiertos de CC son esenciales; y (b) las controversias legales en torno al material con derechos de autor en los datos de entrenamiento. A principios de 2024, la Fundación Mozilla publicó un informe en profundidad (basado en entrevistas con el personal de Common Crawl) titulado *"Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI."* (Source: www.mozillafoundation.org). Este informe reveló estadísticas actualizadas (9.5 PB de datos, 84% de los tokens de GPT-3 provenientes de CC) y proporcionó información actualizada sobre la organización. Casi al mismo tiempo, un notable caso legal (New York Times vs. OpenAI/Microsoft) puso a Common Crawl en los titulares, ya que el contenido del NYT fue extraído en CC y, por lo tanto, utilizado inadvertidamente en GPT-3 (Source: www.mozillafoundation.org). El equipo de Common Crawl también anunció nuevos servicios (por ejemplo, el alojamiento de un *Common Crawl Index* consultable (Source: commoncrawl.org) y una mayor participación de la comunidad (artículos, tutoriales, hackatones).

A lo largo de su historia, Common Crawl se ha mantenido fiel a su misión original de **acceso abierto**. Nunca se transformó en un motor de búsqueda comercial o un proveedor de datos. En cambio, se ha centrado en construir una infraestructura robusta y escalable y una comunidad en torno a los datos abiertos. El liderazgo del proyecto enfatiza regularmente que *"proporcionar datos de entrenamiento para IA nunca fue el propósito principal de Common Crawl,"* y que siempre han acogido a una amplia base de usuarios (siendo los investigadores de IA solo un grupo) (Source: www.96layers.ai). No obstante, como discutiremos, el advenimiento de la IA generativa ha hecho que Common Crawl sea más influyente que nunca, tanto para bien (facilitando la investigación) como para la controversia (preocupaciones sobre derechos de autor y sesgos).

Detalles Técnicos de los Datos de Common Crawl

Formatos de Datos y Acceso

Cada rastreo de Common Crawl produce un conjunto de archivos en formato **WARC** (Web ARChive), que empaqueta secuencias de respuestas HTTP (las páginas web obtenidas) con metadatos. Estos archivos WARC son la salida bruta del rastreo, típicamente nombrados por la fecha y el identificador del rastreo. Además de los WARC, Common Crawl publica una variedad de archivos complementarios:

- **Texto Extraído (archivos WAT):** Para cada WARC, un archivo "WAT" correspondiente contiene metadatos analizados (por ejemplo, encabezados HTTP, enlaces, metadatos JSON).
- **Texto Extraído (archivos WET):** Un archivo "WET" transmite el texto plano extraído de cada página HTML (esencialmente el contenido de texto limpio). Estos permiten a los usuarios analizar rápidamente el texto sin tener que analizar el HTML ellos mismos.
- **Índice de URL (CDX):** Un índice CSV/JSON de todas las URL obtenidas y sus desplazamientos en los WARC, útil para consultar sitios o páginas específicas.
- **Gráficos Web:** Datos de gráficos que enlazan páginas o dominios (por ejemplo, gráficos de enlaces de host a host). Estos se proporcionan periódicamente (por ejemplo, anualmente) para estudiar la conectividad.
- **Tablas de Dominio:** Archivos agregados que listan todos los dominios rastreados y el número de páginas.

Todos estos archivos se almacenan en **cubos de AWS S3** (y se replican en otros lugares). Common Crawl fomenta el uso de análisis en la nube (por ejemplo, Amazon Athena o EMR) para consultar los datos a escala. Por ejemplo, Amazon Athena permite consultas SQL a través del índice de todas las URL o incluso el contenido WARC si está estructurado correctamente. El costo de ejecutar tales consultas es bajo (y a veces cubierto por créditos), lo que lo hace práctico para que los equipos de investigación extraigan conjuntos de datos de Common Crawl sin copiar terabytes a sus servidores locales.

Common Crawl mismo proporciona algunas herramientas para desarrolladores y documentación (por ejemplo, el proyecto “Index to WARC Files and URLs” (Source: registry.opendata.aws). Pero también existe un ecosistema externo vibrante: numerosos proyectos y tutoriales en GitHub (por ejemplo, CC-pyspark, commoncrawljob) ayudan a los nuevos usuarios a empezar. La lista de correo pública de Common Crawl y las comunidades de Slack/Discord están activas con consejos y código compartido.

CCBot (Rastreador de Common Crawl)

El propio **rastreador web**, apodado **CCBot**, se ejecuta continuamente durante cada rastreo mensual. Opera aproximadamente así: un programador maestro despacha instancias de rastreadores (en AWS EC2) que obtienen páginas en paralelo, siguiendo la lista de URL a visitar. Se añaden nuevas URL a la cola a medida que se descubren enlaces. El rastreador utiliza las características estándar de Nutch: respeto por `robots.txt`, limitación automática por dominio y lógica de deduplicación para evitar rastrear interminablemente el mismo contenido (por ejemplo, eliminando parámetros de sesión).

CCBot se identifica con una cadena de agente de usuario, pero Common Crawl recomienda a los webmasters no incluirlo en la lista blanca únicamente por eso, ya que los rastreadores maliciosos pueden falsificarlo (Source: datadome.co). (En su lugar, los propietarios de sitios pueden usar rangos de IP conocidos de AWS para identificar el tráfico de CCBot.) A pesar de ser un usuario legítimo, las direcciones IP de CCBot provienen de grupos dinámicos de AWS, por lo que algunos sitios lo bloquean o limitan inadvertidamente. Common Crawl se esfuerza por ser un rastreador "educado". Por ejemplo, rota rangos de IP, se retira de sitios sobrecargados y permite algunos errores de rastreo. Los administradores de servidores que deseen respetar las normas de la comunidad pueden permitir explícitamente a CCBot ajustando su `robots.txt` (Common Crawl tiene documentación sobre cómo hacerlo).

Con el tiempo, CCBot ha sido refinado para mejorar la eficiencia. La arquitectura actual (a partir de 2025) utiliza un sistema distribuido y tolerante a fallos en AWS, coordinado por el equipo central (dirigido por un "ingeniero de rastreo"). El rastreo de mayo de 2025, por ejemplo, cubrió **2.47 mil millones de páginas** (ver informe de la cumbre de Twitter (Source: commoncrawl.org). En total, el sistema ha demostrado ser **escalable**: Common Crawl señala con orgullo que su rastreo es ahora *“gargantuesco”*, mucho más allá de la capacidad de cualquier investigador académico para duplicarlo (Source: nonprofitquarterly.org).

Pipeline de Procesamiento de Datos

Las páginas rastreadas en bruto pasan por un pipeline de procesamiento antes de su publicación. Los pasos clave incluyen:

- **Extracción de Enlaces:** Identificar todos los hipervínculos en cada página para añadir a la frontera de rastreo. Construir gráficos de enlaces (a nivel de dominio y de host) para el análisis.
- **Deduplicación de Contenido:** Filtrar páginas idénticas o casi idénticas para reducir el desperdicio y el sesgo. Common Crawl aplica una deduplicación agresiva a nivel de documento y página para que los datos archivados tengan una redundancia mínima.
- **Extracción de Texto:** Eliminar HTML/CSS y extraer el contenido de texto, que se almacena en los archivos “WET”. Esto incluye la detección de idioma (Common Crawl típicamente se enfoca en texto en inglés, pero también capturará otros idiomas).
- **Metadatos HTTP:** Registrar los encabezados de respuesta, el tipo de contenido y la información del servidor para cada obtención (en los archivos WAT).
- **Manejo de Errores:** Registrar cualquier error de obtención o tiempo de espera en un archivo de "erratas". Common Crawl mantiene **un registro de erratas** que lista URL o dominios que fallan consistentemente, para mejorar futuros rastreos.

El resultado final es un producto de datos rico: para cualquier mes dado, un usuario puede recuperar no solo los blobs HTML en bruto, sino también un corpus paralelo de oraciones (el texto WET) y toda la estructura de hipervínculos. El código del pipeline es de código abierto, y las mejoras (por ejemplo, mejor análisis de HTML, manejo de JavaScript) se integran periódicamente.

(En febrero de 2023, Common Crawl anunció en su blog que tenía la intención de experimentar con el *pre-renderizado* de páginas que requieren JavaScript, pero a finales de 2025, el corpus principal sigue centrado en HTML.)

Características del Conjunto de Datos

- **Distribución de Idiomas:** Los menús de Common Crawl revelan que el conjunto de datos es multilingüe, pero fuertemente sesgado hacia el inglés. Según el informe de Mozilla, el rastreo es **"principalmente en inglés"** con una cobertura regional variable. Por ejemplo, se han derivado de CC conjuntos de datos de 50 millones de artículos de noticias en alemán (Source: commoncrawl.org) y otros corpus específicos de idiomas, pero el rastreo en bruto tiene mucho más contenido en inglés.
- **Diversidad de Sitios:** Common Crawl intenta equilibrar amplitud y profundidad. Incluye sitios importantes (noticias, comercio electrónico, blogs) así como sitios web de cola larga. Sin embargo, no apunta a la "web profunda" ni a páginas protegidas con contraseña. Tampoco puede rastrear sitios que prohíben los bots o requieren inicios de sesión.
- **Instantáneas Temporales:** Cada rastreo mensual tiene una marca de tiempo. En consecuencia, los archivos de Common Crawl pueden usarse para estudiar la evolución de la web (por ejemplo, cómo una página o dominio cambia con el tiempo). Sin embargo, Common Crawl *no* es un archivo continuo como Wayback Machine; no conserva todas las versiones de una página diariamente; principalmente proporciona una "toma" por URL al mes (a menos que la página cambie y se vuelva a rastrear más tarde).

En conjunto, los datos de Common Crawl son extremadamente grandes y bastante representativos de la web pública (sujetos a bots y acceso). Es *el* archivo web más grande disponible públicamente para uso de investigación, combinando volumen con accesibilidad.

Casos de Uso e Impacto

El conjunto de datos abierto de Common Crawl ha permitido una enorme variedad de aplicaciones. Organizamos su uso en varias categorías amplias:

1. IA y Aprendizaje Automático (LLMs, Embeddings, etc.)

Common Crawl se ha convertido en **la fuente de datos fundamental para el procesamiento del lenguaje natural a gran escala y la IA**. Prácticamente todos los modelos de lenguaje modernos han recurrido a estos datos. Por ejemplo:

- **GPT-3 y ChatGPT:** Cuando OpenAI entrenó GPT-3 (que subyace a ChatGPT), la mayoría de sus tokens de entrenamiento provino de Common Crawl. El artículo publicado por OpenAI sobre GPT-3 muestra que *"la mayor cantidad de datos de entrenamiento proviene de Common Crawl"* (Source: datadome.co). Un análisis de Mozilla corrobora esto: encontró que *más del 80% de los tokens de GPT-3* se originaron en Common Crawl (Source: www.mozillafoundation.org). (Las GPU típicamente se entrenan con múltiples corpus; para GPT-3, las otras fuentes fueron WebText2, libros y Wikipedia. Pero Common Crawl fue la porción más grande.) Debido a que GPT-3 alimenta directamente a los chatbots y asistentes de IA, el contenido de Common Crawl (bueno o malo) esencialmente "habla" a los usuarios finales a través de la IA.

- **Otros Grandes Modelos de Lenguaje:** Muchos otros LLM notables se construyeron con datos de CC:

  - Los modelos T5 y basados en BERT de Google incorporaron subconjuntos de Common Crawl.
  - RoBERTa de Facebook fue entrenado con una mezcla de datos de CC y noticias en 2019.
  - Modelos de código abierto como GPT-NeoX de EleutherAI y modelos más pequeños como GPT-2 utilizaron CC.
  - El modelo **Grover** (2019) de Zellers *et al.* – un modelo para generar y detectar noticias falsas – utilizó explícitamente Common Crawl para el texto web (Source: dallascard.github.io).
  - Más recientemente, la mayoría de los nuevos modelos (Bellatrix, LLaMA, etc.) utilizan pipelines como The Pile o RefinedWeb, que a su vez se extraen de instantáneas de Common Crawl (Source: dallascard.github.io). De hecho, las instantáneas de Common Crawl se reempaquetan en conjuntos de datos derivados (por ejemplo, C4, Colossal Clean Crawls) que alimentan cargas de trabajo de entrenamiento a gran escala.
  - Una encuesta de 47 LLM diversos (2019-2023) encontró que *"al menos el 64%"* de ellos fueron entrenados con datos de Common Crawl (Source: www.mozillafoundation.org). Esto incluye modelos de nueva generación como ChatGPT-4 (a través

de GPT-4), LLaMA de Meta, Mistral, Claude 2, etc. (Algunos modelos también pueden usar datos propietarios o mixtos, pero CC sigue siendo un pilar fundamental.)

- **Embeddings de Palabras y Herramientas de PNL:** El conjunto de datos ha permitido recursos fundamentales de PNL. Los clásicos **embeddings GloVe** (840B tokens, inglés) y **embeddings FastText** (600B tokens) se entrenan con texto de CC (Source: huggingface.co). Corpus de código abierto como Colossal Clean Crawls (C4) y conjuntos de datos multilingües derivados de Common Crawl impulsan modelos de traducción y resumidores. La investigación en modelado de temas, análisis de sentimientos, recuperación de información y más a menudo utiliza CC como fuente de texto en bruto. Por ejemplo, un estudio de 2019 construyó un corpus paralelo bilingüe a partir de CC para la traducción automática (Source: huggingface.co).

- **Chatbots y Asistentes de IA:** Más allá del entrenamiento de modelos offline, algunos servicios realizan rastreos en tiempo real de CC para apoyar la IA. Por ejemplo, **DeepSeek** y algunas plataformas de búsqueda "impulsadas por IA" ingieren páginas de CC para proporcionar sus respuestas. Muchos bots de IA también dependen de CC para verificar hechos o aumentar las respuestas, ya que es un índice conveniente de la web pública.

- **Datos para Modelos de Visión y Multimodales:** Si bien Common Crawl contiene principalmente texto, también incluye URL de imágenes (y ocasionalmente metadatos de imágenes). Empresas como **TinEye** aprovechan el índice de URL de imágenes de CC para construir servicios de búsqueda inversa de imágenes (Source: nonprofitquarterly.org). (TinEye utilizó explícitamente Common Crawl para encontrar imágenes similares a una imagen de consulta.) Algunos modelos de visión de IA utilizan subtítulos de texto alineados con CC o texto alternativo en los datos de CC para emparejar con imágenes.

En resumen, **investigadores y empresas de IA utilizan intensamente Common Crawl** como fuente de datos gratuita. Su ubicuidad en el entrenamiento de modelos ha planteado tanto oportunidades (avanzar en la IA) como preocupaciones (sesgos, derechos de autor), más sobre esto a continuación.

2. Investigación Académica y Científica

El corpus de Common Crawl es ampliamente citado en la investigación académica, en diversas disciplinas:

- **Lenguaje Natural y Ciencia Web:** Los investigadores analizan el uso y los patrones del lenguaje. Por ejemplo, CC se ha utilizado para estudiar la estructura de hipervínculos (quién enlaza a quién en la web), geolocalizar noticias (se construyó un conjunto de datos de 50 millones de artículos de noticias en alemán a partir de CC (Source: commoncrawl.org), y analizar la legibilidad o frases comunes en la web. El trabajo sobre **gráficos web** (teoría de grafos aplicada a dominios) a menudo utiliza los datos de gráficos de enlaces de CC (Source: commoncrawl.org).

- **Minería de Datos y Análisis de Big Data:** El conjunto de datos ejemplifica los "grandes datos abiertos". Los investigadores prueban algoritmos de minería de texto a gran escala (agrupación, detección de valores atípicos, análisis de temas) en CC. La capacidad de acceder a petabytes de datos del mundo real ha permitido estudios comparativos de pipelines de procesamiento de texto.

- **Estudios de Recuperación de Información (IR):** Common Crawl se utiliza para construir motores de búsqueda experimentales. Por ejemplo, **Elastic ChatNoir** en Bauhaus Weimar está construido para buscar en los archivos de ClueWeb y Common Crawl (Source: commoncrawl.org). Los investigadores de IR también evalúan algoritmos de clasificación en subconjuntos de CC, o usan CC como referencia para el contenido de páginas web. El propio equipo de Common Crawl proporciona una API de "Búsqueda Rápida Simple" (CCSS) para búsquedas rápidas de palabras clave sobre el índice.

- **Ciberseguridad y Medición de Abusos:** La naturaleza a gran escala de CC permite escanear patrones maliciosos. Por ejemplo, el artículo “Lurking Malice in the Cloud” (ACM 2016) escaneó todas las páginas de CC para encontrar scripts incrustados vinculados a dominios de malware conocidos (Source: huggingface.co). Los investigadores han utilizado CC para cuantificar la prevalencia de encabezados HTTP (in)seguros, bibliotecas desactualizadas o scripts de cryptojacking en sitios web populares.

- **Economía y Ciencias Sociales:** Los científicos sociales utilizan CC como un proxy para el discurso público. Por ejemplo, un estudio utilizó CC para analizar la moderación y censura de contenido; la investigación *Citizen Lab* “Banned Books” analizó páginas de productos de Amazon extraídas a través de CC para detectar políticas de censura (Source: commoncrawl.org). Otros

casos de uso incluyen el seguimiento de la desinformación sobre salud, el análisis de la propaganda política o el estudio de la difusión de contenido en múltiples idiomas en la web abierta.

- **Índices de Citas y Cartografía de la Ciencia:** La disponibilidad de miles de millones de citas académicas obtenidas de textos de CC ha permitido incluso la metainvestigación. Por ejemplo, la repetición de análisis de citas y la construcción de grafos de conocimiento a escala colosal.

En particular, el propio sitio web de Common Crawl destaca muchos trabajos de investigación: selecciona enlaces a trabajos publicados que aprovechan los datos de CC (Source: commoncrawl.org). Las citas abarcan NeurIPS/ICLR para PNL, conferencias WWW/WWW para análisis web, y revistas de IA, ciencia de la información y ciencias sociales computacionales.

3. Aplicaciones comerciales e industriales

Más allá del ámbito académico, numerosas empresas y startups han construido productos sobre los datos de Common Crawl. Algunos ejemplos notables:

- **Búsqueda de imágenes – TinEye:** Como se mencionó, **TinEye** (de Idée Inc.) utiliza Common Crawl para indexar imágenes. Cuando un usuario envía una imagen, TinEye le aplica un hash y busca en los datos de imágenes recopilados de CC para encontrar otras similares (Source: nonprofitquarterly.org). CC proporcionó una fuente grande y gratuita de imágenes y sus URL, lo que permitió a TinEye lanzar un negocio viable sin tener que rastrear la web por sí mismos.

- **Análisis de impacto – Lucky Oyster:** Lucky Oyster Labs (adquirida por Rendever) utilizó Common Crawl para la escucha social y el análisis de tendencias. Construyeron herramientas sobre CC para "dar sentido a lo que la gente discute en la web" como un motor de información (Source: nonprofitquarterly.org). (El artículo de NPQ menciona a Lucky Oyster como una startup que aprovecha CC, aunque los detalles son ahora escasos).

- **Búsqueda como servicio – Caso Crate.IO:** Algunas empresas desarrollaron conectores y motores para consultar datos de CC. Por ejemplo, Crate.IO publicó un blog sobre "importación desde fuentes de datos personalizadas" utilizando un plugin, mostrando cómo alimentar los archivos de CC en su base de datos SQL (Source: commoncrawl.org). Del mismo modo, "CommonCrawlJob" y "CommonCrawlScalaTools" son proyectos de GitHub que ayudan a cargar datos de CC en sistemas de big data. Estos son en su mayoría pruebas de concepto o herramientas para desarrolladores.

- **Motores de búsqueda de startups:** Al menos un equipo emprendedor (Elastic ChatNoir (Source: commoncrawl.org) construyó una interfaz de motor de búsqueda específicamente para clones de Common Crawl del conjunto de datos ClueWeb. Otro, las instantáneas web abiertas de Carrot Search, han experimentado con CC. Existe interés en crear motores de búsqueda sin fines de lucro o alternativos utilizando CC como backend de datos, evitando la necesidad de rastrear la web por sí mismos.

- **Marketing y SEO:** Algunas empresas de análisis SEO utilizan CC para estimar el acceso al sitio o el análisis de la competencia. Aunque la mayoría de los productos SEO comerciales se basan en rastreadores propietarios, CC ofrece un conjunto de datos gratuito para evaluar el número global de páginas o las tendencias de contenido. Por ejemplo, las líneas de código para herramientas SEO como Majestic o Ahrefs podrían incorporar datos de CC para el análisis de backlinks, aunque los detalles suelen ser propietarios.

- **Publicidad e inteligencia empresarial:** Las empresas de datos (incluida Factual, la empresa fundada por Gil Elbaz) han integrado datos de CC para enriquecer los conjuntos de datos empresariales. Por ejemplo, el recuento de dominios, la frescura del sitio y la clasificación de contenido se pueden obtener de CC para alimentar la segmentación de anuncios o las herramientas de marketing B2B. Sin embargo, debido a la naturaleza automatizada de los datos, los conocimientos basados en CC deben validarse cuidadosamente para uso comercial.

La Tabla 2 (a continuación) resume algunos casos de uso y proyectos ilustrativos que aprovechan los datos de Common Crawl:

USUARIO/PROYECTO	CASO DE USO	FUENTE / NOTAS
TinEye	Búsqueda inversa de imágenes (encontrar imágenes similares rastreando)	Utiliza imágenes rastreadas por CC (Source: nonprofitquarterly.org). (IDée Inc.)
Lucky Oyster	Análisis de tendencias sociales/culturales	Startup que utiliza CC para analizar tendencias de contenido web (Source: nonprofitquarterly.org).
GloVe (Stanford)	Embeddings de vectores de palabras (840B tokens de CC)	CC proporcionó texto para el modelo GloVe (Source: huggingface.co).
GPT-3/ChatGPT	Datos de entrenamiento para modelos de lenguaje grandes (~80% de tokens de CC)	Informe de Mozilla: *"Más del 80% de los tokens de GPT-3 provienen de Common Crawl."* (Source: www.mozillafoundation.org).
Modelos de lenguaje	Entrenamiento/ajuste fino (RoBERTa, T5, LLaMA, etc.)	Los LLM (2019–2023) a menudo utilizan corpus basados en CC (Source: dallascard.github.io) (Source: www.mozillafoundation.org).
Motores de búsqueda	Construcción de índices de búsqueda alternativos (ej. ChatNoir)	Elastic ChatNoir: búsqueda de datos de CC (Source: commoncrawl.org). (Bauhaus-Weimar)
Investigación en PNL	Análisis estadístico de texto web (modelos de temas, resumen)	Decenas de artículos académicos en dominios de PNL citan a CC.
Métricas web	Estudios de censura/libertad de expresión (ej. censura de Amazon)	Citizen Lab "Banned Books" utilizó CC (Source: commoncrawl.org); otros artículos de ciencia web.

(Tabla 2: Ejemplos seleccionados de cómo se utilizan los datos de Common Crawl en la práctica, con citas.)

Además de estos ejemplos, el propio sitio web de Common Crawl enumera **numerosos proyectos**: conjuntos de datos abiertos (WikiSQL de tablas web), experimentos de búsqueda basados en la nube, tutoriales de Elasticsearch y cursos académicos, todos construidos sobre datos de CC (Source: commoncrawl.org). Anécdoticamente, Gil Elbaz ha comentado que **"si no eres Google, OpenAI o Microsoft, casi todo el mundo depende de Common Crawl"** para datos a gran escala (Source: www.96layers.ai). Esto subraya cuán omnipresente se ha vuelto CC para cualquier organización que no pueda implementar su propio rastreador web a la escala de Google.

Casos de estudio

Para ilustrar el impacto de Common Crawl de manera más concreta, describimos dos estudios de caso detallados: uno sobre IA/entrenamiento de modelos y otro sobre búsqueda abierta.

Caso de estudio: GPT-3 y la revolución de los LLM

Como un ejemplo de alto perfil, consideremos GPT-3 de OpenAI (2020) y sus modelos hermanos. Estos "Transformers Generativos Preentrenados" logran impresionantes habilidades de lenguaje natural, pero su poder se deriva de vastos datos de entrenamiento. Common Crawl desempeñó un papel estelar:

- **Composición del conjunto de datos:** El artículo de GPT-3 (Brown et al. 2020) enumera las fuentes de datos: WebText2 (el propio rastreo de OpenAI de páginas vinculadas a Reddit), Google Books, Wikipedia y **Common Crawl**. En tamaño bruto, Common Crawl fue, con mucho, el más grande. Un análisis posterior confirma que *"la mayor cantidad de datos de*

entrenamiento proviene de Common Crawl" (Source: datadome.co). El informe de Mozilla aclara que **más del 80% de todos los tokens utilizados por GPT-3 provenían de CC** (Source: www.mozillafoundation.org).

- **Modelo resultante:** GPT-3-175B, con 175 mil millones de parámetros, fue entrenado con 570 GB de datos de texto filtrados (alrededor de 500 mil millones de tokens). Si el 80% provino de CC, eso significa ~456 GB de texto de CC. Esta escala sería imposible sin un corpus web existente. La disponibilidad de CC significó que OpenAI no necesitó asignar recursos para rastrear la web por sí mismos en ese momento (aunque probablemente también tenían algunos datos web internos).

- **Uso profesional:** Cuando se lanzó GPT-3, se integró rápidamente en productos (por ejemplo, Copilot de Microsoft, ChatGPT de OpenAI en 2022). Estos servicios actúan entonces como una "capa de IA" sobre CC. Algunos usuarios se preocupan de que, a medida que ChatGPT genera respuestas, pueda regurgitar texto de páginas de Common Crawl sin atribución. De hecho, el informe de Mozilla señala que los modelos basados en CC a menudo producen *contenido sesgado o con derechos de autor* porque tienden a memorizar los datos de entrenamiento.

- **Implicaciones legales (Caso NYT):** A finales de 2023, The New York Times demandó a OpenAI, alegando que los datos de entrenamiento de ChatGPT (GPT-3.5/GPT-4) incluían indebidamente contenido del Times. Common Crawl se convirtió en una pieza clave de evidencia porque los artículos del Times habían sido extraídos a CC antes de que se entrenara el modelo, y OpenAI utilizó esas instantáneas de CC. Una hoja informativa de Mozilla explica: *"El contenido del NYT constituía una proporción significativa de los datos de Common Crawl en el momento en que OpenAI lanzó ChatGPT, y por lo tanto, probablemente constituyó una porción significativa de los datos de entrenamiento de GPT-3"* (Source: www.mozillafoundation.org). Esto destaca cómo la apertura de CC puede conducir inadvertidamente a una exposición legal cuando el texto con derechos de autor se redistribuye en modelos.

- **Diversidad y sesgo:** Debido a que tantos LLM dependen de CC, las directrices aprendidas en CC se propagan ampliamente. Si CC carece de suficiente contenido de ciertos idiomas o datos demográficos, los modelos pueden tener un rendimiento inferior en esos temas. La investigación de Mozilla advierte que *"el conjunto de datos de Common Crawl incluye deliberadamente contenido problemático (toxicidad, discurso de odio, etc.) para apoyar la investigación sobre esos fenómenos."* Por el contrario, muchas pipelines de entrenamiento de IA filtran CC intensamente (por ejemplo, solo conservan "páginas en inglés de alta calidad") (Source: www.mozillafoundation.org), lo que significa que la toxicidad bruta de CC puede influir en el comportamiento del modelo si no se elimina cuidadosamente.

En resumen, el caso de GPT-3 muestra que **Common Crawl se ha convertido en la columna vertebral de la investigación de IA generativa** en la década de 2020. Redujo drásticamente la barrera para entrenar modelos grandes. El hecho de que los datos de una pequeña organización sin fines de lucro estén impulsando sistemas de IA multimillonarios es notable. También obliga a una reflexión: cuando un conjunto de datos abierto alimenta una IA de código cerrado, ¿quién asume la responsabilidad por el contenido? La dirección de Common Crawl enfatiza que los datos estaban destinados a todo tipo de análisis (incluida la investigación sobre el discurso de odio), no explícitamente para entrenar modelos de miles de millones de dólares (Source: www.96layers.ai). El debate de la comunidad ahora gira en torno a cómo asegurar que los modelos basados en CC sean "confiables" (eliminando sesgos, respetando los derechos de autor, etc.) (Source: www.mozillafoundation.org).

Caso de estudio: Búsqueda abierta a través de Common Crawl

Otro caso ilustrativo son los intentos de **construir motores de búsqueda utilizando datos de Common Crawl**. Si bien empresas expertas en la web como Google o Bing desarrollan sus propios rastreadores, algunos grupos independientes han explorado el uso de CC como fuente de datos para servicios de búsqueda alternativos.

- **Elastic ChatNoir:** Investigadores de la Universidad Bauhaus crearon *ChatNoir*, una interfaz de búsqueda abierta para los corpus ClueWeb y CC (Source: commoncrawl.org). Esto está dirigido a la investigación en humanidades digitales y recuperación de información. ChatNoir indexa páginas de Common Crawl y proporciona una interfaz de búsqueda simple, permitiendo a los usuarios consultar el archivo de CC como si fuera un motor de búsqueda. Esto demuestra que, en principio, se puede usar CC como el "backend" para la búsqueda.

- **CC Search (Beta):** El propio Common Crawl lanzó **CC Search** (ahora operado por el equipo de Creative Commons/WordPress) que permite a los usuarios buscar CC por palabras clave. El sitio web de CC menciona actualizaciones como *"Grandes cambios para CC Search Beta"* a finales de 2024 (escrito por Paola Villarrela). El objetivo es hacer que los datos de CC sean más accesibles (por ejemplo, añadiendo búsqueda por licencia, idioma, etc.).

- **Propuestas de startups:** La idea de un "motor de búsqueda sin fines de lucro" ha sido planteada periódicamente (incluso en Hacker News (Source: dallascard.github.io). Incluso el titular del artículo de Nonprofit Quarterly fue "Conozca a Common Crawl, la organización sin fines de lucro que podría remodelar la web" (Source: nonprofitquarterly.org). Por ahora, Common Crawl en sí mismo sigue siendo solo datos (sin portal de búsqueda para usuarios), pero terceros pueden construir sobre él. La existencia de CC significa que cualquier grupo con suficientes recursos podría lanzar un motor de búsqueda *sin rastrear la web por sí mismos*.

- **Consideraciones prácticas:** Es importante señalar que los datos de Common Crawl tienen limitaciones para la búsqueda: no incluyen PageRank, datos de clics de usuarios o frescura actualizada más allá de la granularidad mensual. Algunos sitios web excluyen a CC, y el conjunto de datos está "congelado" en puntos mensuales. Por lo tanto, un motor basado en CC estaría parcialmente desactualizado. Sin embargo, proyectos de búsqueda "específicos de dominio" a pequeña escala han utilizado CC con éxito. Por ejemplo, un equipo de investigación podría restringir CC a dominios de noticias y construir una búsqueda de noticias especializada.

En el comercio electrónico o SEO, algunas empresas extraen CC para recopilar información abierta sobre datos de productos o clasificaciones de sitios. Se informa que un blogger (Claus Matzinger de Crate.IO) escribió sobre la importación de datos de CC a una base de datos compatible con la búsqueda (Source: commoncrawl.org). Como dijo un observador de CC desde hace mucho tiempo: *"Si no eres Google, OpenAI o Microsoft... casi todo el mundo depende de Common Crawl"* (Source: www.96layers.ai) para al menos algunos datos a gran escala.

Estos casos demuestran que Common Crawl ha permitido **nuevos tipos de servicios** que antes solo los gigantes de la búsqueda podían contemplar. Si bien ningún motor de búsqueda comercial importante (con consultas en vivo) ha adoptado completamente CC, el proyecto ha reducido efectivamente la barrera: construir un sistema de búsqueda experimental o académico sobre Common Crawl es sencillo y rentable.

Análisis de datos y hallazgos de investigación

Más allá de las anécdotas de uso, los investigadores han analizado cuantitativamente el propio Common Crawl. Algunos hallazgos representativos:

- **Escala de datos:** Una entrevista de 2024 con Stefan Baack, investigador de Mozilla, resumió el volumen mensual e histórico de Common Crawl (Source: www.96layers.ai). Por ejemplo, señala que cada archivo mensual tiene 90 TB comprimidos y Common Crawl ha acumulado *"más de 250 mil millones de páginas web"* en 17 años (Source: www.96layers.ai). Estas cifras son consistentes con la afirmación del sitio oficial de *"más de 300 mil millones de páginas"* (Source: commoncrawl.org). Tal análisis subraya el tamaño inigualable de CC.

- **Métricas de citas:** Al rastrear Google Scholar o bases de datos bibliográficas, el personal de Common Crawl descubrió que sus datos tenían más de 10,000 citas en la literatura académica (Source: commoncrawl.org) (Source: dallascard.github.io). Esto demuestra la amplia adopción en diversos campos. Los investigadores han indicado que CC se utiliza en campos tan variados como la detección de spam web, bibliotecas digitales, periodismo (seguimiento de noticias falsas) e incluso informática de la salud (por ejemplo, escaneo de desinformación médica).

- **Cobertura de idioma y sitio:** El informe de Mozilla destaca que **el inglés domina** Common Crawl. Muestra el recuento de páginas web por país/idioma, y señala que muchas páginas chinas, japonesas y de redes sociales (por ejemplo, Facebook, Twitter) faltan o están subrepresentadas debido a restricciones de rastreo (Source: www.mozillafoundation.org). De hecho, las páginas de sitios que bloquean explícitamente los rastreadores están ausentes. El informe también señala que el objetivo de CC de apoyar la "investigación sobre el discurso de odio" significa que incluye dicho contenido intencionalmente (Source: www.mozillafoundation.org), lo cual es una elección de diseño (se deja sin filtrar para permitir el análisis). Sin embargo, aquellos interesados en el entrenamiento de LLM a menudo *filtran* estas páginas.

- **Robustez Técnica:** Se ha realizado un análisis de los datos de registro de CC para evaluar el propio rastreo web. Por ejemplo, el artículo de Springer "Web Crawl Refusals: Insights from Common Crawl" estudió cómo los servidores web bloquean o limitan a los rastreadores, utilizando los propios registros de obtención de CC (Source: commoncrawl.org). Los resultados sirvieron para informar sobre las mejores prácticas de rastreo (por ejemplo, cómo lidiar con los bloqueos falsos de "fake chatgpt-bot").

- **Riqueza Semántica de los Datos:** Algunos proyectos han intentado anotar CC a escala. Por ejemplo, creando grafos de conocimiento extrayendo entidades y relaciones del texto de CC. El proyecto CSRankings de Stanford utiliza CC para medir el tamaño de las publicaciones de CS de CVPR, ICML, NeurIPS (aunque esto es una digresión). Pero más relevante: los investigadores han utilizado CC para construir grafos de conocimiento de "sentido común" abiertos analizando miles de millones de oraciones.

En resumen, el **metaanálisis** de Common Crawl confirma su escala e influencia. Estudios independientes han validado las estadísticas brutas del sitio y han explorado sus sesgos. Dichos estudios retroalimentan la mejora del conjunto de datos (por ejemplo, destacando regiones de la web subrastreadas) y guían a los usuarios sobre el uso apropiado (por ejemplo, advierte sobre problemas de derechos de autor) (Source: www.mozillafoundation.org).

Desafíos, Limitaciones y Problemas

Aunque los datos de Common Crawl son potentes, no están exentos de desafíos o críticas:

- **Sesgo y Representatividad:** Como se ha señalado, CC está sesgado en cuanto a idioma (principalmente inglés) y región (más EE. UU./UE). Algunos campos (como el contenido africano y asiático) están subrepresentados. Esto puede sesgar cualquier análisis o IA entrenada con CC. El informe de Mozilla advierte explícitamente que CC *no* debe ser tratado como un "sustituto de toda la web" (Source: www.mozillafoundation.org). Los investigadores a menudo complementan CC con otros corpus para una mejor cobertura (por ejemplo, noticias, archivos gubernamentales, colecciones específicas de idiomas).

- **Calidad del Contenido:** Common Crawl incluye deliberadamente una amplia variedad de contenido, lo que significa que también captura *páginas web de baja calidad, spam o tóxicas*. No hay un filtrado estricto de contenido "bueno" frente a "malo" por defecto. Para algunos casos de uso (investigación lingüística, detección de sesgos), esta inclusividad es una característica. Pero para el entrenamiento de IA, requiere una limpieza adicional. Por ejemplo, el inteligente artículo de Ablestacks sobre The Pile y conjuntos de datos similares incluye múltiples filtros para eliminar blasfemias, contenido para adultos, texto no inglés, etc. El análisis de Mozilla enfatiza que los creadores de IA deben "eliminar" el contenido no deseado de CC si su objetivo es un entrenamiento de modelos seguro (Source: www.mozillafoundation.org). En la práctica, muchas tuberías de IA (Aleph, Redwood, etc.) utilizan listas de origen colectivo o heurísticas para filtrar CC.

- **Derechos de Autor y Licencias:** Los "Términos de Uso" de CC establecen que las páginas web se recopilan sin tener en cuenta los derechos de autor, asumiendo que el texto en la web pública puede ser utilizado (similar al funcionamiento de Googlebot). Sin embargo, el auge de la IA ha planteado problemas legales. La mencionada demanda del New York Times sugiere que CC pudo haber extraído miles de artículos con derechos de autor en sitios de noticias, y que luego estos terminaron en los parámetros de GPT-3. Esto ilustra una tensión: Common Crawl cree que su recopilación de datos está legalmente protegida (por ejemplo, bajo las excepciones de la DMCA para el almacenamiento en caché/rastreo, o bajo la idea de "uso transformador" en el entrenamiento de IA). Pero los titulares de derechos no están de acuerdo. Common Crawl no pidió permiso específicamente a cada creador de contenido en la web; se basa fundamentalmente en los términos de servicio de Internet y robots.txt. A finales de 2023, Common Crawl aclaró que una vez que el contenido está en CC, está "ahí para que todos lo usen (lo que incluye el ajuste fino y la inferencia / aumento por recuperación)" (Source: www.mozillafoundation.org). Esta postura es controvertida.

- **Comité y Gobernanza:** Dado que CC es gestionado por voluntarios, su futuro depende de la buena voluntad continua y el apoyo de los patrocinadores. No hay financiación garantizada ni una gran dotación. Si los principales donantes tecnológicos retiraran su apoyo, las operaciones de CC podrían verse comprometidas. Sin embargo, a partir de 2025, el interés en conservar los proyectos de datos web abiertos parece alto, dado el interés legislativo en la regulación de la IA y la ciencia abierta. Common Crawl tiene planes (según las últimas declaraciones) para diversificar la financiación y posiblemente añadir nuevas características (como metadatos de licencias, APIs de exclusión, etc.) para abordar las preocupaciones de los propietarios de contenido.

- **Limitaciones Técnicas:** El conjunto de datos es masivo, pero aún puede omitir contenido generado dinámicamente o oculto detrás de formularios. Los sitios que utilizan una representación pesada del lado del cliente o que requieren JavaScript pueden ser parcialmente invisibles para los rastreadores de solo texto. Algunas páginas modernas (por ejemplo, aplicaciones de una sola página) con poco HTML estático podrían no ser capturadas correctamente. Common Crawl ha experimentado con

navegadores sin interfaz gráfica, pero esto es costoso. Por lo tanto, CC puede subindexar sitios muy modernos y con mucho JavaScript. Además, debido a que realiza una pasada al mes, puede omitir actualizaciones rápidas o páginas efímeras. Los usuarios que necesitan datos frescos en tiempo real no pueden depender únicamente de CC.

En general, el equipo de Common Crawl reconoce estos desafíos. Su estrategia ha sido la transparencia: publican con frecuencia entradas de blog y respuestas para explicar el alcance y los límites del conjunto de datos (por ejemplo, "Web Archiving File Formats Explained" (Source: commoncrawl.org). Animan a los usuarios a ver CC como una **infraestructura compartida**, similar a un experimento abierto, en lugar de un producto perfeccionado.

Direcciones Futuras e Implicaciones

Mirando hacia el futuro, Common Crawl se encuentra en la intersección de varias tendencias en la ciencia de datos y la gobernanza de Internet:

- **Escalado de la Calidad de los Datos:** Common Crawl podría adoptar un filtrado o etiquetado más avanzado para servir mejor a los usuarios. Por ejemplo, generar un subconjunto "limpio" del rastreo (eliminando contenido probablemente spam o para adultos) podría ayudar a la adopción generalizada. Por el contrario, crear subrastreos especializados (por ejemplo, un rastreo multilingüe o un rastreo en inglés de alta calidad) podría atraer a nuevas audiencias.
- **Propietarios de Contenido y Permisos:** A medida que evolucionan los debates sobre los derechos de los datos, Common Crawl podría implementar mecanismos de exclusión. Ya, algunos sitios ofrecen reglas `DDD/Robots.txt` para la exclusión de IA. Common Crawl se ofreció a respetar las `x-robot-tags` que bloquean todo el rastreo no bot (estilo DRM). Los sistemas futuros podrían permitir a los propietarios de sitios solicitar la eliminación del archivo de CC. Por otro lado, tales exclusiones amenazan la uniformidad de los conjuntos de datos para los investigadores. Es probable que el proyecto continúe colaborando con expertos legales para lograr un equilibrio.
- **Iniciativas de Búsqueda Abierta:** Existe una creciente defensa de la "infraestructura de búsqueda como un servicio público". Common Crawl podría convertirse en la base de datos de una nueva generación de motores de búsqueda abiertos o grafos de conocimiento. Por ejemplo, proyectos como **OpenWebIndex** (un proyecto propuesto financiado por la UE) hacen eco de la misión de Common Crawl. Podríamos ver asociaciones donde el rastreo de Common Crawl impulse índices especializados (por ejemplo, un motor de búsqueda académico de contenido educativo o una búsqueda de compras abierta). El lanzamiento de la **API de Índice** de Common Crawl (anunciada en 2023) muestra un movimiento en esta dirección.
- **IA y Uso Responsable:** Dado que los datos de Common Crawl alimentan la IA generativa, la fundación podría invertir en características de "ética de la IA". Esto podría incluir anotaciones (marcando páginas que son propaganda o desinformación de salud) o la integración de filtros de despolarización. El informe de Mozilla sugiere que los constructores deben añadir "filtros de datos robustos" (Source: www.mozillafoundation.org); Common Crawl mismo podría comenzar a ofrecer versiones prefiltradas o herramientas para el filtrado (por ejemplo, un filtro de toxicidad).
- **Análisis Adicional por Common Crawl:** La fundación podría producir más análisis de datos internamente. Por ejemplo, su GitHub muestra paneles de "Crawl Stats" y "Graph Stats". Expandirlos para mostrar desgloses de idiomas en tiempo real, métricas de diversidad de dominios o tendencias semánticas podría ser valioso. Esto ayudaría tanto a los usuarios como a los financiadores a comprender el alcance del recurso.
- **Asociaciones Globales:** Para mejorar la cobertura, Common Crawl podría asociarse con universidades u ONG internacionales para sembrar el rastreo con más contenido global (por ejemplo, a través de los 100 dominios principales específicos de cada país). También podría colaborar con bibliotecas nacionales (como Europeana o archivos web nacionales) para integrar jardines vallados de la web.

En términos más amplios, el impacto de Common Crawl sugiere que los **bienes comunes de datos** (infraestructura de datos abiertos) podrían ser un modelo viable para otros dominios: imagine corpus abiertos de artículos científicos, imágenes o sensores ambientales. El éxito de Common Crawl proporciona una plantilla: equipo mínimo, patrocinadores en la nube, datos abiertos. Demuestra que, bajo las condiciones adecuadas, *"los datos son la nueva infraestructura pública"*.

Conclusión

Common Crawl surgió de la visión de Gil Elbaz de un índice web abierto, y durante casi dos décadas se ha convertido en un recurso fundamental para la innovación basada en datos. Su **historia** es una historia de comienzos modestos (una pequeña organización sin fines de lucro en 2007) que escaló a través del esfuerzo comunitario y el apoyo de la nube para convertirse en un **archivo web gigantesco** (Source: nonprofitquarterly.org) (Source: www.96layers.ai). Nació de un compromiso con los datos abiertos y se ha adherido a ese principio: hacer que la información a escala web sea democráticamente accesible, no propietaria.

Hoy en día, Common Crawl es utilizado por miles de investigadores y desarrolladores en todo el mundo. Impulsa la vanguardia de la IA (prácticamente todos los grandes modelos de lenguaje dependen de él) y permite a las nuevas empresas que de otro modo no podrían permitirse la infraestructura de Google. La Tabla 2 de este informe ilustró algunos ejemplos concretos, pero un recuento exhaustivo sería aún más largo. Su presencia en más de **10.000** publicaciones académicas (Source: commoncrawl.org) (Source: dallascard.github.io) es un testimonio de su influencia.

Sin embargo, un gran poder conlleva grandes responsabilidades y complicaciones. El uso de Common Crawl en el entrenamiento de IA ha planteado problemas sociales y legales, especialmente a medida que los modelos generativos dan forma al discurso público. El equipo de Common Crawl es consciente de ello y ha colaborado con la comunidad sobre cómo utilizar los datos de forma responsable. El informe de Mozilla y otros análisis sugieren que CC será parte de los debates sobre la ética de la IA y los derechos de autor durante años (Source: www.mozillafoundation.org) (Source: www.mozillafoundation.org).

Mirando hacia adelante, la trayectoria de Common Crawl parece encaminada a una expansión continua y una integración más profunda con la investigación abierta. A medida que el poder computacional crece y la IA busca cada vez más datos, el valor del archivo web abierto de Common Crawl probablemente aumentará. La comunidad que lo rodea podría expandirse, quizás pasando de un pequeño equipo a un consorcio colaborativo más grande. Hay proyectos incipientes para extender sus capacidades (como índices de búsqueda más ricos u opciones de filtrado) que podrían dar forma a la era de la "Búsqueda 2.0" (Source: commoncrawl.org).

En resumen, la **historia completa** de Common Crawl es un caso de estudio sobre cómo una iniciativa pequeña y bien dirigida puede **abrir drásticamente los bienes comunes de datos**. Comenzó como una respuesta a los temores de monopolio en la búsqueda web, y de hecho ha abierto puertas a la innovación. Su fundador Gil Elbaz y sus colaboradores lograron crear "la web como una base de datos gigante", accesible para todos (Source: nonprofitquarterly.org). La historia de Common Crawl, desde el primer rastreo de cinco mil millones de páginas hasta los miles de millones de páginas actuales, ilustra el poder de la infraestructura abierta. Su papel futuro probablemente se profundizará a medida que la sociedad aborde los beneficios y desafíos de la IA a escala web y la ciencia abierta.

Todas las afirmaciones anteriores están respaldadas por fuentes citadas de la propia documentación de Common Crawl, informes de medios, entrevistas y análisis académicos (Source: commoncrawl.org) (Source: www.96layers.ai) (Source: www.mozillafoundation.org) (Source: dallascard.github.io). Estas referencias proporcionan una base transparente para las afirmaciones del informe. Al cubrir múltiples perspectivas (técnica, organizacional, ética) e incluir datos cuantitativos (recuentos de páginas, estadísticas de uso, citas reportadas), hemos buscado una **cuenta exhaustiva y basada en evidencia** de la historia, el estado actual y las implicaciones futuras de Common Crawl.

Tags: common-crawl, rastreo-web, datos-entrenamiento-llm, datos-abiertos, gil-elbaz, big-data, repositorio-web, apache-nutch

---