

Forbes Vetted Best Product Awards

Our editors tested hundreds of products to name 150 standouts of the year.

Read Now

INNOVATION > AI & BIG DATA


Common Crawl And Unlocking Web Archives For Research

By [Kalev Leetaru](#), Contributor. ⓘ I write about the broad intersectio... 

Follow Author

Sep 28, 2017, 10:22pm EDT

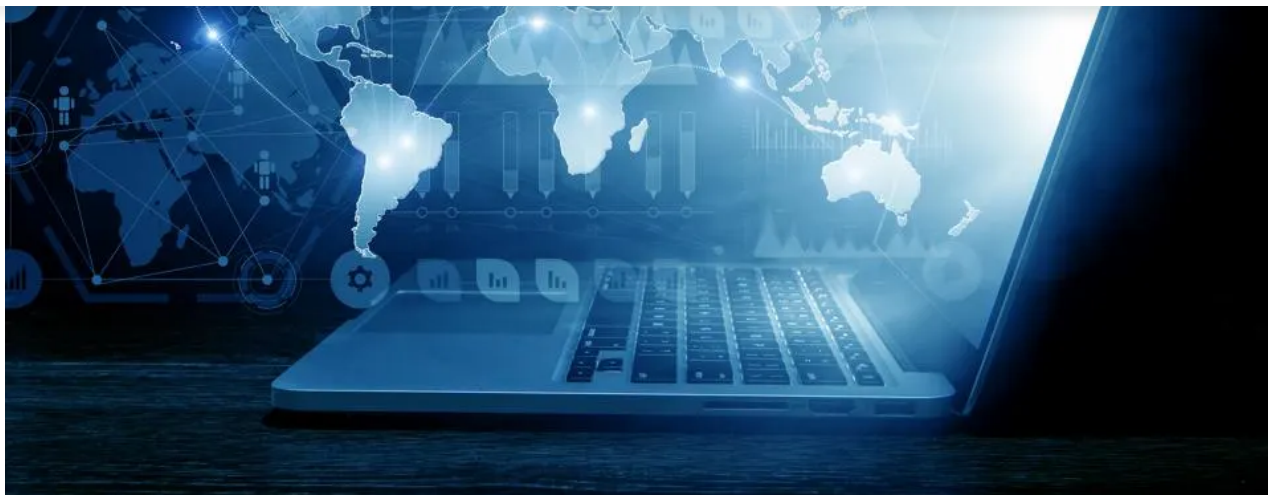
 Share  Save

 This article is more than 7 years old.

Forbes Vetted Best Product Awards

Our editors tested hundreds of products to name 150 standouts of the year.

Read Now



Shutterstock

Over the past half-decade I've written extensively about web archiving, including [why](#) we need to [understand](#) what's in our massive [archives](#) of the [web](#), whether our archives are failing to capture the [modern](#) and [social](#) web, the need for archives to [modernize](#) their technology infrastructures and, perhaps most intriguingly for the world of "big data," how archives can make their [petabytes](#) of holdings [available for research](#). What might it look like if the world's web archives opened up their collections for academic research, making hundreds of billions of web objects totaling tens of petabytes and stretching back to the founding of the modern web available as a massive shared corpus to power the modern data mining revolution, from studies of the evolution of the web to powering the vast training corpuses required to build today's cutting edge neural networks?

When it comes to crawling the open web to build large corpuses for data mining, universities in the US and Canada have largely adopted a [hands-off](#) approach, exempting most work from ethical review, granting permission to ignore terms of use or copyright restrictions and waiving traditional policies on data management and replication on the grounds that material harvested from the open web is publicly accessible information and that its copyright owners, by virtue of being making it available on the web without password protection, [encourage](#) its access and use.

On the other hand, the world's non-profit and governmental web archives, whom

collectively hold tens of petabytes of archived content crawled from the open web stretching back 20+ years, have as a whole largely resisted opening their collections

Forbes Vetted Best Product Awards

Our editors tested hundreds of products to name 150 standouts of the year.

Read Now

While some archives have cited technical limitations in making their content more accessible, the most common argument against offering bulk data mining access revolves around copyright law and concern that by boxing up gigabytes, terabytes or even petabytes of web content and redistributing it to researchers, web archives could potentially be viewed as “redistributing” copyrighted content. Given the growing interest among large content holders in licensing their material for precisely such bulk data mining efforts, some archives have expressed concern that traditional application of “fair use” doctrine in potentially permitting such data mining access may be gradually eroding.

Thus, paradoxically, research universities have largely adopted the stance that researchers are free to crawl the web and bulk download vast quantities of content to use in their data mining research, while web archives as a whole have adopted the stance that they cannot make their holdings available for data mining because they would, in their view, be “redistributing” the content they downloaded to third parties to use for data mining.

One large web archive has bucked this trend and stood alone among its peers: [Common Crawl](#). Similar to other large web archiving initiatives like the Internet Archive, Common Crawl conducts regular web wide crawls of the open web and preserves all of the content it downloads in the standard WARC file format. Unlike many other archives, it focuses primarily on preserving HTML web pages and does not archive images, videos, JavaScript files, CSS stylesheets, etc. Its goal is not to preserve the exact look and feel of a website on a given snapshot in time, but rather to collect a vast cross section of HTML web pages from across the web in a single place to enable large-scale data mining at web scale.

Yet, what makes Common Crawl so unique is that it makes everything it crawls freely available for download for research. Each month it conducts an open web crawl, boxes up all of the HTML pages it downloads and makes a set of WARC files and a few derivative file formats available for download.

Its most recent crawl, covering August 2017, contains more than 3.28 billion pages totaling 280TiB, while the previous month's crawl contains 3.16 billion pages and

Forbes Vetted Best Product Awards

Our editors tested hundreds of products to name 150 standouts of the year.

Read Now

HTML content saved by preservation-focused web archives like the Internet Archive, this vast compilation of web pages cannot be used to reproduce a page's appearance as it stood on a given point in time. Instead, it is primarily useful for large-scale data mining research, exploring questions like the linking structure of the web or analyzing the textual content of pages, rather than acting as a historical replay service.

The project excludes sites which have robots.txt exclusion policies, following the historical policy of many other web archives, though it is worth noting that the Internet Archive earlier this year began slowly [phasing out](#) its reliance on such files due to their detrimental effect on preservation completeness. Common Crawl also allows sites to request removal from their index. Other than these cases, Common Crawl attempts to crawl as much of the remaining web as possible, aiming for a representative sample of the open web.

Moreover, Common Crawl has made its data publicly available for more than half a decade and has become a staple of large academic studies of the web with high visibility in the research community, suggesting that its approach to copyright compliance and research access appears to be working for it.

Yet, beyond its [summary](#) and [full](#) terms of use documents, the project has published little in terms of how it views its work fitting into US and international standards on copyright and fair use, so I reached out Sara Crouse, Director of Common Crawl, to speak to how the project approaches copyright and fair use and any advice they might have for other web archives considering broadening access to their holdings for academic big data research.

Ms. Crouse noted the risk adverse nature of the web archiving community as a whole (historically many adhered and still adhere to a strict "[opt in](#)" policy requiring prior [approval](#) before crawling a site) and the unwillingness of many archives to modernize their thinking on copyright and to engage more closely with the legal community in ways that could help them expand fair use horizons. In particular, she

noted “since we [in the US] are beholden to the Copyright Act, while living in a digital age, many well-intentioned organizations devoted to web science, archiving,

Forbes Vetted Best Product Awards

Our editors tested hundreds of products to name 150 standouts of the year.

[Read Now](#)

Given that US universities as a whole have moved aggressively towards this idea of expanding the boundaries of fair use and permitting opt-out bulk crawling of the web to compile research datasets, Common Crawl seems to be in good company when it comes to interpreting fair use for the digital age and modern views on utilizing the web for research.

Returning to the difference between Common Crawl’s datasets and traditional preservation-focused web archiving, Ms. Crouse emphasized that they capture only HTML pages and exclude multimedia content like images, video and other dynamic content.

She noted that a key aspect of their approach to fair use is that web pages are intended for consumption by human beings one at a time using a web browser, while Common Crawl concatenates billions of pages together in the specialized WARC file format designed for machine data mining. Specifically, “Common Crawl does not offer separate/individual web pages for easy consumption. The three data formats that are provided include text, metadata, and raw data, and the data is concatenated” and “the format of the output is not a downloaded web page. The output is in WARC file format which contains the components of a page that are beneficial to machine-level analysis and make for space- efficient archiving (essentially: header, text, and some metadata).”

In the eyes of Common Crawl, the use of specialized archival-oriented file formats like WARC (which is the format of choice of most web archives) limit the content’s use to transformative purposes like data mining and, combined with the lack of capture of styling, image and other visual content, renders the captured pages unsuitable to human browsing, transforming them from their originally intended purpose of human consumption.

As Ms. Crouse put it, “this is big data intended for machine learning/readability. Further, our intention for its use is for public benefit i.e. to encourage research and innovation, not direct consumption.” She noted that “from the layperson’s

perspective, it is not at all trivial at present to extract a specific website's content (that is, text) from a Common Crawl dataset. This task generally requires one to

Forbes Vetted Best Product Awards

Our editors tested hundreds of products to name 150 standouts of the year.

Read Now

research sectors. At a higher level, it's important to note that we provide a broad and representative sample of the web, in the form of web crawl data, each month. No one really knows how big the web is, and at present, we limit our monthly data publication to approximately 3 billion pages.”

Of course, given that content owners are increasingly looking to bulk data mining access licensing as a revenue stream, this raises the concern that even if web archives are transforming content designed for human consumption into machine friendly streams designed for data mining, such transformation may conflict with copyright holders' own bulk licensing ambitions. For example, many of the large content licensors like LexisNexis, Factiva and Bloomberg all offer licensed commercial bulk feeds designed to support data mining access that pay royalty fees to content owners for their material that is used.

Common Crawl believes it addresses this through the fact that its archive represents only a sample of each website crawled, rather than striving for 100% coverage. Specifically, Ms. Crouse noted that “at present, [crawls are] in monthly increments that are discontinuous month-to-month. We do only what is reasonable, necessary, and economical to achieve a representative sample. For instance, we limit the number of pages crawled from any given domain so, for large content owners, it is highly probable that their content, if included in a certain month's crawl data, is not wholly represented and thus not ideal for mining for comprehensive results ... if the content owner is not a large site, or in a niche market, their URL is less likely to be included in the seeds in the frontier, and, since we limit depth (# of links followed) for the sake of both economy and broader representative web coverage, 'niche' content may not even appear in a given month's dataset.”

To put it another way, Common Crawl's mission is to create a “representative sample” of the web at large by crawling a sampling of pages and limiting the number of pages from each site they capture. Thus, their capture of any given site will represent a discontinuous sampling of pages that can change from month to month. A researcher wishing to analyze a single web site in its entirety would therefore not

be able to turn to Common Crawl and would instead have to conduct their own crawl of the site or turn to a commercial aggregator that partners with the content holder

Forbes Vetted Best Product Awards

Our editors tested hundreds of products to name 150 standouts of the year.

[Read Now](#)

sample” of the web at large, rather than attempting to capture a single site in its entirety (and in fact ensuring that it does not include more than a certain number of pages per site), the crawl self-limits itself to being applicable only to macro-level research examining web scale questions. Such “web scale” questions cannot be answered through any existing open dataset and by incorporating specific design features Common Crawl ensures that more traditional research questions, like data mining the entirety of a single site, which might be viewed as redistribution of that site or competing with its owner’s ability to license its content for data mining, is simply not possible.

Thus, to summarize, Common Crawl is both similar to other web archives in its workflow of crawling the web and archiving what it finds, but sets itself apart by focusing on creating a representative sample of HTML pages from across the entire web, rather than trying to preserve the entirety of a specific set of websites with an eye towards visual and functional preservation. Even when a given page is contained in Common Crawl’s archives, the technical sophistication and effort required to extract it and the lack of supporting CSS, JavaScript and image/video files renders the capture useless for the kind of non-technical browser-based access and interaction such pages are designed for.

Of course, copyright and what counts as “fair use” is a notoriously complex, contradictory, contested and ever-changing field and only time will tell whether Common Crawl’s interpretation of fair use holds up and becomes a standard that other web archives follow. At the very least, however, Common Crawl presents a powerful and intriguing model for how web-scale data can power open data research and offers traditional web archives a set of workflows, rationales and precedent to examine that are fully aligned with those of the academic community. Given its popularity and continued growth over the past decade it is clear that Common Crawl’s model is working and that many of its underlying approaches are highly applicable to the broader web archiving community.

Putting this all together, today’s web archives preserve for future generations the

dawn of our digital society, but lock those tens of petabytes of documentary holdings away in dark archives or permit only a page at a time to be accessed. Common

Forbes Vetted Best Product Awards

Our editors tested hundreds of products to name 150 standouts of the year.

[Read Now](#)

archives will follow Common Crawl's example and explore ways of shaping the future of fair use and gradually opening their doors to research, all while ensuring that copyright and the rights of content holders are respected.

[Editorial Standards](#)

[Forbes Accolades](#)

Forbes

© 2025 Forbes Media LLC. All Rights Reserved.

[AdChoices](#) [Privacy Statement](#) [Do Not Sell or Share My Personal Information](#)
[Limit the Use of My Sensitive Personal Information](#) [Privacy Preferences](#)
[Digital Terms of Sale](#) [Terms of Service](#) [Contact Us](#) [Send Us Feedback](#)
[Report a Security Issue](#) [Jobs At Forbes](#) [Reprints & Permissions](#) [Forbes Press Room](#)
[Advertise](#)