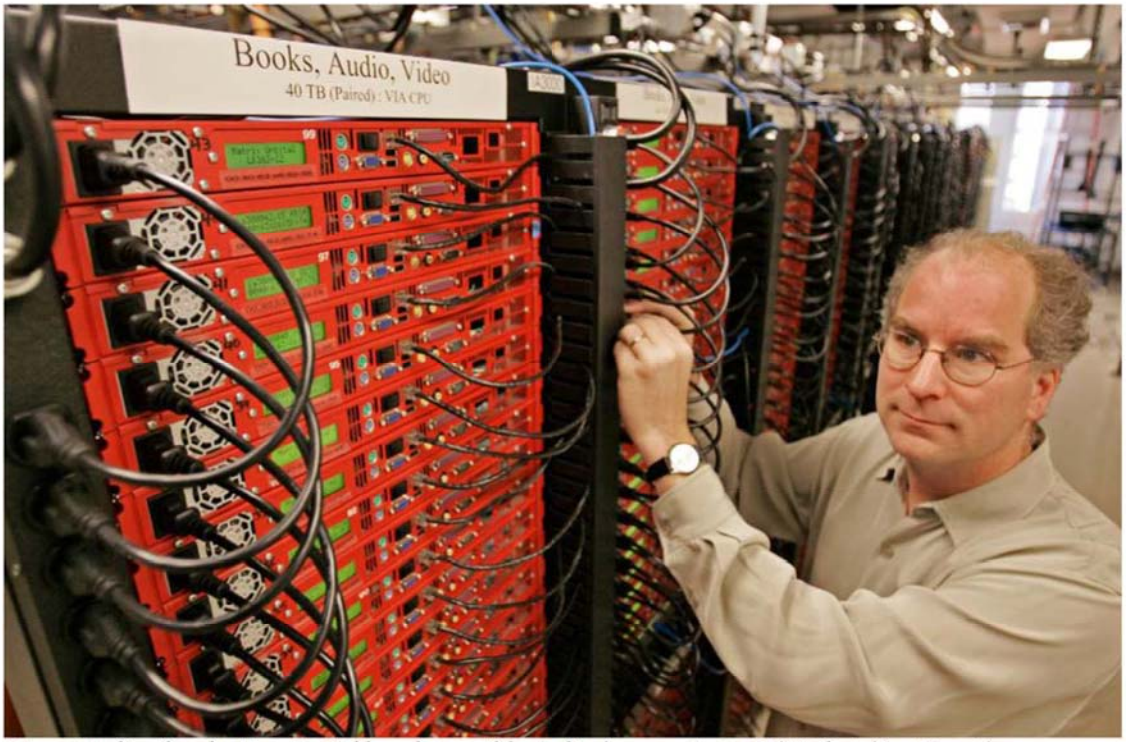
Forbes / Tech
JAN 18, 2016 @ 10:59 AM 4,256

The Internet Archive Turns 20: A Behind The Scenes Look At Archiving The Web

Kalev Leetaru, CONTRIBUTOR
I write about the broad intersection of data and society. FULL BIO

Opinions expressed by Forbes Contributors are their own.



Internet Archive founder Brewster Kahle and some of the Archive's servers in 2006. (AP Photo/Ben Margot)

To most of the web surfing public, the Internet Archive's Wayback Machine is the face of the Archive's web archiving activities. Via a simple interface, anyone can type in a URL and see how it has changed over the last 20 years. Yet, behind that simple search box lies an exquisitely complex assemblage of datasets and partners that make possible the Archive's vast repository of the web. How does the Archive really work, what does its crawl workflow look like, how does it handle issues like robots.txt, and what can all of this teach us about the future of web archiving?

Perhaps the first and most important detail to understand about the Internet Archive's web crawling activities is that it operates far more like a traditional library archive than a modern commercial search engine. Most large web crawling operations today operate vast farms of standardized crawlers all operating in unison, sharing a common set of rules and behaviors. They traditionally operate in continuous crawling mode, in which the goal is to scour the web 24/7/365 and attempt to identify and ingest every available URL.

In contrast, the Internet Archive is comprised of a myriad independent datasets, feeds and crawls, each of which has very different characteristics and rules governing its construction, with some run by the Archive and others by its many partners and contributors. In the place of a single standardized continuous crawl with stable criteria and algorithms, there is a vibrant collage of inputs that all feed into the Archive's sum holdings. As Mark Graham, Director of the Wayback Machine put in an email, the Internet Archive's web materials are comprised of "many different collections driven by many organizations that have different approaches to crawling." At the time of this writing, the primary web holdings of the Archive total more than 4.1 million items across 7,357 distinct collections, while its Archive-It program has over 440 partner organizations overseeing specific targeted collections. Contributors range from middle school students in Battle Ground, WA to the National Library of France.

Those 4.1 million items comprise a treasure trove covering nearly every imaginable topic and data type. There are crawls contributed by the Sloan Foundation and Alexa, crawls run by IA on behalf of NARA and the Internet Memory Foundation, mirrors of Common Crawl and even DNS inventories containing more than 2.5 billion records from 2013. Many specialty archives preserve the final snapshots of now-defunct online communities like GeoCities and Wretch. Dedicated Archive-It crawls preserve a myriad hand-selected or sponsored websites on an ongoing basis such as the Wake Forest University Archives. These dedicated Archive-IT crawls can be accessed directly and in some cases appear to feed into the Wayback Machine, accounting for why the Wake Forest site is captured almost every Thursday and Friday over the last two years like clockwork.

Alexa Internet has been a major source of the Archive's regular crawl data since 1996, with the Archive's FAQ page stating "much of our archived web data comes from our own crawls or from Alexa Internet's crawls ... Internet Archive's crawls tend to find sites that are well linked from other sites ... Alexa Internet uses its own methods to discover sites to crawl. It may be helpful to install the free Alexa toolbar and visit the site you want crawled to make sure they know about it."

Another prominent source is the Archive's "Worldwide Web Crawls," which are described as "Since September 10th, 2010, the Internet Archive has been running Worldwide Web Crawls of the global web, capturing web elements, pages, sites and parts of sites. Each Worldwide Web Crawl was initiated from one or more lists of URLs that are known as 'Seed Lists' ... various rules are also applied to the logic of each crawl. Those rules define things like the depth the crawler will try to reach for each host (website) it finds." With respect to how frequently the Archive crawls each site, the only available insight is "For the most part a given host will only be captured once per Worldwide Web Crawl, however it might be captured more frequently (e.g. once per hour for various news sites) via other crawls."

The most recent crawl appears to be Wide Crawl Number 13, created on January 9, 2015 and running through present. Few details are available regarding the crawls, though the March 2011 crawl (Wide 2) states it ran from March 9, 2011 to December 23, 2011, capturing 2.7 billion snapshots of 2.3 billion unique URLs from a total of 29 million unique websites. The documentation notes that it used the Alexa Top 1 Million ranking as its seed list and excluded sites with robots.txt directives. As a warning for researchers, the collection notes "We also included repeated crawls of some Argentinian government sites, so looking at results by country will be somewhat skewed."

Augmenting these efforts, the Archive's No More 404 program provides live feeds from the GDELT Project, Wikipedia and WordPress. The GDELT Project provides a daily list of all URLs of online news coverage it monitors around the world, which the Archive then crawls and archives, vastly expanding the Archive's reach into the non-Western world. The Wikipedia feed monitors the "[W]ikipedia IRC channel for updated article[s], extracts newly added citations, and feed[s] those URLs for crawling," while the WordPress feed scans "WordPress's official blog update stream, and schedules each permalink URL of new post for crawling." These greatly expand the Archive's holdings of news and other material relating to current events.

Some crawls are designed to make a single one-time capture to ensure that at least one copy of everything on a given site is preserved, while others are designed to intensively recrawl a small subset of hand-selected sites on a regular interval to ensure both that new content is found and that all previously-identified content is checked for any changes and freshly archived. In terms of how frequently the Archive recrawls a given site Mr. Graham wrote that "it is a function of the hows, whats and whys of our crawls. The Internet Archive does not crawl all sites equally nor is our crawl frequency strictly a function of how popular a site is." He goes on to caution "I would expect any researcher would be remiss to not take the fluid nature of the web, and the crawls of the [Internet Archive], into consideration" with respect to interpreting the highly variable nature of the Archive's recrawl rate.

Though it acts as the general public's primary gateway to the Archive's web materials, the Wayback Machine is merely a public interface to a limited subset of all these holdings. Only a portion of what the Archive crawls or receives from external organizations and partners is made available in the Wayback Machine, though as Mr. Graham noted there is at present "no master flowchart of the source of captures that are available via the Wayback Machine" so it is difficult to know what percent of the holdings above can be found through the Wayback Machine's public interface. Moreover, large portions of the Archive's holdings carry notices that access to them is restricted, often due to embargos, license agreements, or other processes and policies of the Archive.

In this way, the Archive is essentially a massive global collage of crawls and datasets, some conducted by the Archive itself, others contributed by partners. Some focus on the open web, some focus on the foundations of the web's infrastructure, and others focus on very narrow slices of the web as defined by contributing sponsors or Archive staff. Some are obtained through donations, some through targeted acquisitions, and others compiled by the Archive itself, much in the way a traditional paper archive operates. Indeed, the Archive is even more similar to traditional archives in its use of a dark archive in which only a portion of its holdings are publically accessible, with the rest having various access restrictions and documentation ranging from detailed descriptions to simple item placeholders.

This is in marked contrast to the description that is often portrayed of the Archive by outsiders as a traditional centralized continuous crawl infrastructure, with a large farm of standardized crawlers ingesting the open web and feeding the Wayback Machine akin to what a traditional commercial search engine might do. The Archive has essentially taken the traditional model of a library archive and brought it into the digital era, rather than take the model of a search engine and add a preservation component to it.

There are likely many reasons for this architectural decision. It is certainly not the difficulty of building such systems – there are numerous open source infrastructures and technologies that make it highly tractable to build continuous web-scale crawlers given the amount of hardware available to the Archive. Indeed, I myself have been building global web scale crawling systems since 1995 and while still a senior in high school in 2000 launched a whole-of-web continuous crawling system with sideband recrawlers and an array of realtime content analysis and web mining algorithms running at the NSF-supported supercomputing center NCSA.

Why then has the Archive employed such a patchwork approach to web archival, rather than the established centralized and standardized model of its commercial peers? Part of this may go back to the Archive's roots. When the Internet Archive was first formed Alexa Internet was the primary source of its collections, donating its daily open crawl data. The Archive therefore had little need to run its own whole-of-web crawls, since it had a large commercial partner providing it such a feed. It could instead focus on supplementing that general feed with specialized crawls focusing on particular verticals and partner with other crawling organizations to mirror their archives.

From the chronology of datasets that make up its web holdings, the Archive appears to have evolved in this way as a central repository and custodian of web data, taking on the role of archivist and curator, rather than trying to build its own centralized continuous crawl of the entire web. Over time it appears to have taken on an ever-expanding collection role of its own, running its own general purpose web-scale crawls and bolstering them with a rapidly growing assortment of specialized crawls.

With all of this data pouring in from across the world, a key question is how the Internet Archive deals with exclusions, especially the ubiquitous "robots.txt" crawler exclusion protocol.

The Internet Archive's Archive-It program appears to strictly enforce robots.txt files, requiring special permission for a given crawl to ignore them: "By default, the Archive-It crawler honors and respects all robots.txt exclusion requests. On a case by case basis institutions can set up rules to ignore robots.txt blocks for specific sites, but this is not available in Archive-It accounts by default. If you think you may need to ignore robots.txt for a site, please contact the Archive-It team for more information or to enable this feature for your account."

In contrast, the Library of Congress uses a strict opt-in process and "notifies each site that we would like to include in the archive (with the exception of government websites), prior to archiving. In some cases, the e-mail asks permission to archive or to provide off-site access to researchers." The Library uses the Internet Archive to perform its crawling and ignores robots.txt for those crawls: "The Library of Congress has contracted with the Internet Archive to collect content from websites at regular intervals ... the Internet Archive uses the Heritrix crawler to collect websites on behalf of the Library of Congress. Our crawler is instructed to bypass robots.txt in order to obtain the most complete and accurate representation of websites such as yours." In this case, the Library views the written archival permission as taking precedent over robots.txt directives: "The Library notifies site owners before crawling which means we generally ignore robots.txt exclusions."

The British Library appears to ignore robots.txt in order to preserve page rendering elements and for selected content deemed culturally important, stating "Do you respect robots.txt? As a rule, yes: we do follow the robots exclusion protocol.

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE