

Most Popular

1. He found forgotten letters from the '70s in his attic. Turns out they...
2. How 49ers' Trey Lance and other first-round QBs stack up
3. Here's why California has the lowest COVID rate in the nation
4. Acc tear pas

TECH

He founded the Internet Archive with a utopian vision. That hasn't changed, but the internet has



Chase DiFelicantonio

Updated: Sep. 6, 2021 6:28 p.m.



Internet Archive founder Brewster Kahle shows statues of the organization's staff.

Constanza Hevia H./Special to The Chronicle

Inside his library, Brewster Kahle is dancing. He smiles as he sways on the spot, an antique Victrola filling the foyer of the building, a former church, with the scratchy jazz tunes of yesteryear.

He lifts the needle and the music stops, but just for now. Soon his staff will convert the aging record to a string of ones and zeroes that will live forever in cyberspace. This is the Internet Archive, and that is why Kahle, and it, are here: To make available for free, online, every bit of digital or physical information that exists.

To walk with Kahle through his columned temple of knowledge in San Francisco's Richmond District is to understand the scale of what he and his staff, which now numbers more than 100, have been hard at work for almost 25 years. In a loading area stacks of donated books await their turn on a specialized scanning machine where, shrouded behind a black curtain, a technician painstakingly copies endless pages.

Downstairs, microfiche reels are being converted into computer images that will join the staggering amount of data the archive has collected over the years.

Its servers holds more than 70 unique petabytes of data — 70 million gigabytes — including 65 million texts, movies, audio files, images, books and more.

Kahle's quest to build what he calls "A Library of Alexandria for the Internet" started in the 1990s when he began sending out programs called crawlers to take digital snapshots of every page on the web, hundreds of billions of which are available to anyone through the archive's Wayback Machine.

That vision of free and open access to information is deeply entwined with the early ideals of Silicon Valley and the origins of the Internet itself.

“The reason for the internet and specifically the World Wide Web was to make it so that everyone’s a publisher and everybody can go and have a voice,” Kahle said. To him, the need for a new type of library for that new publishing system, the internet, was obvious.

But while’s Kahle’s aims have not changed, the internet has. That early Utopian vision of the positive forces of digital interconnectedness is increasingly at odds with the troves of copyrighted and paywalled material online that grows every day.



Left: An Albany (N.Y.) Times newspaper from 1947 at the Internet Archive offices. Right: Book scanner Eliza Zhang opens a box with Albany Times newspapers.

Photos by Constanza Hevia H./Special to The Chronicle

When the archive began its collection, most people online were accessing a few main homepages such as Yahoo.com, said Margaret O’Mara, a professor at the University of Washington and historian of Silicon Valley.

“Now, not only is there so much more information, but also a lot of that information is proprietary,” O’Mara said. “There are questions about how the internet works and how the internet economy works that can’t be answered by capturing web pages or capturing documents or digitizing a magazine.”

Despite that, she said the archive is an invaluable resource for researchers like herself and reflects the idealism at the root of Silicon Valley’s dream of a more

open, connected, and accessible world.

“They are conserving the past in a way that is kind of a rare thing to see in the industry and a community that’s always so focused on the future and focused on what the next thing is,” O’Mara said.

That changing online landscape is on Kahle’s mind as he makes his way into the beating heart of the archive’s cavernous main room. The space is quiet. Diffused with a golden light that filters in through the windows, the former church nave still feels somehow holy. Few people are in the building because of the pandemic, but this room is never really empty, its pews peopled with miniature statues of employees and volunteers past and present, including a bespectacled one of Kahle himself.

Here, the server banks hum and flash with every upload and download as Kahle discusses how libraries, even in cyberspace, can burn.

Across the auditorium flanking the main stage where hymn numbers were once posted, three numbers are picked out in metal: 200, 404 and 451. The first two are common internet codes for when a page is successfully accessed or not. The third shows up when content has been taken down for legal reasons, like copyright infringement.

It is also not coincidentally a reference to Ray Bradbury’s anti-censorship novel “Fahrenheit 451.”



Book scanner Eliza Zhang, one of more than 100 employees, works at the Internet Archive offices in the Richmond District.

Photos by Constanza Hevia H. / Special to The Chronicle

Kahle said in the past that if one library and its books burned, copies probably lived on in another physical space. “That’s not the case on the web,” he said. For example “If a newspaper goes offline in Turkey, all of their archives go. And that’s not the way you can run a culture.”

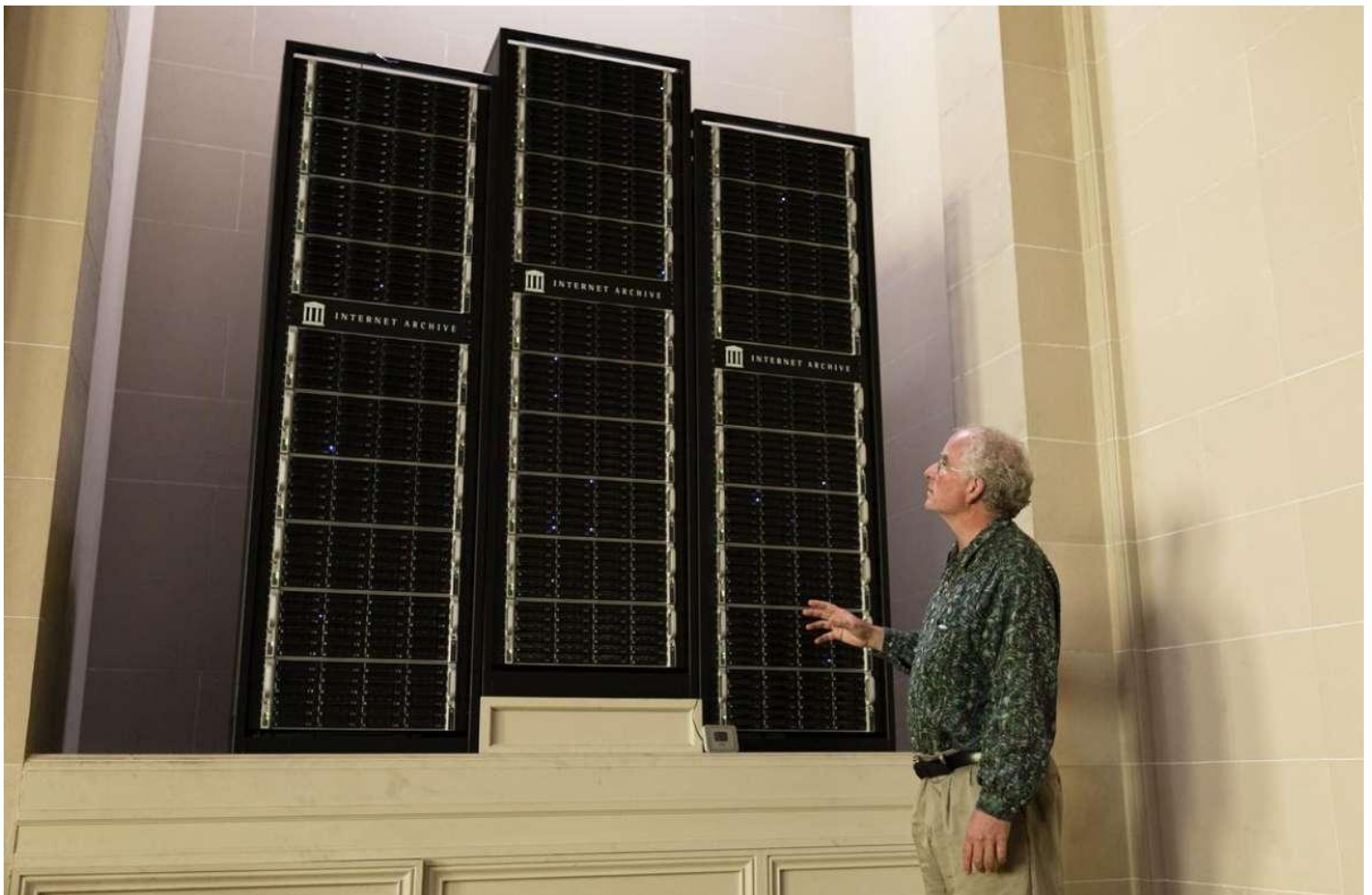
The archive has for years purchased and digitized books, lending them out through its site for free with a wait list like other libraries. But when the coronavirus pandemic hit last year and libraries and schools closed down, the archive created what it called the National Emergency Library, a collection of 1.4 million online books available to users without a wait.

A lawsuit filed by four of the nation's largest publishing houses soon followed, one of the many challenges the archive faces in its quest for freedom of navigation rights in cyberspace.

Kahle maintains that copyright laws don't bar libraries like his from owning, digitizing and lending books out with certain controls in place.

Perhaps an even larger barrier in Kahle's mind are smartphones, and the proprietary and protected apps that fill them.

"These things are full of apps that are not open," He said, holding up his phone on a recent Zoom call. That also means many of them are immune to his crawlers and cannot be saved for posterity. That is a deeply vexing problem for the archive's mission, along with paywalls, which can and do block Kahle's crawlers.



Brewster Kahle, who founded the Internet Archive 25 years ago, discusses the San Francisco organization's servers, which hold more than 70 million gigabytes of data — including 65 million texts, movies, audio files, images, books and more.
Constanza Revla H. / Special to The Chronicle

The original internet format of hypertext links still in use today allows people to “weave knowledge together,” he said. But “app world is innately siloed into corporate products. That’s not the way we’re going to build a culture that inter-operates, builds on each other and can build new ideas.”

Kahle’s career in technology stretches back to the early 1980s when he graduated from the Massachusetts Institute of Technology, where he studied artificial intelligence before graduating. He helped found a supercomputer company called Thinking Machines before creating the internet’s first publishing system called Wide Area Information Server, which was eventually sold to America Online.

Air Quality Tracker



Check levels down to the neighborhood

In the past Kahle also found ways to make money off of software without sacrificing the ideal of the archive. When he sold Alexa Internet, a web research and information company he co-founded in the 1990s, to Amazon, he made a deal with then-CEO Jeff Bezos. He'd sell the software only if Bezos would allow it to keep donating a copy of the Internet to his archive every day. Bezos agreed.

The Internet Archive, today, is funded by many small donations that average around \$20 apiece, according to Katie Barrett, the archive's senior development manager. The archive also makes money scanning books for libraries and receives funding from the Kahle/Austin Foundation Foundation, which was founded with Kahle's wife, Mary Austin.

Tax forms from 2019 show the archive's revenue topping \$36 million for the year, with almost \$30 million of that in contributions and grants.

In its pursuit of a more open and accessible world, the nonprofit works with Wikipedia, fixing links and updating pages that link back to sites that would be lost if the Wayback Machine had not saved them in the first place. Working with the archive, Wikipedia has added more than 25 million archived web pages, mostly from Wayback Machine links, to 150 Wikipedia language editions.

"We share a vision of the Internet where nonprofit services can increase humanity's access to knowledge," Gwadamirai Majange, a spokesperson for the Wikimedia Foundation, which owns Wikipedia, said in an email.



— The Internet Archive building in the Richmond District.
Constanza Hevia H./Special to The Chronicle

The archive also has partnered with groups such as the Digital Public Library of America, contributing mostly digitized print material to its site.

Groups such as the Long Now Foundation also seek to foster that kind of long-term thinking through its 10,000 Year Clock and a project to create a digital library of human language for future generations, partly as a counterpoint to the short-term, profit-driven models of modern tech companies.

Kalhe has also expanded his nonprofit efforts outside of the digital world.

Among those was an ill-fated attempt to establish a credit union with \$1 million from the archive. A more successful bid saw him set up another nonprofit and buy

below-market rates.

For his part Kahle said he recognizes the increasing challenges to the mission, but that hasn't stopped him yet. "I wake up on different sides of the bed saying, you know, this is going to work, and we're making it go," he said. "And then other times it's, like, there's just so much arrayed against us."

Despite that, Kahle's servers continue to blink blue with life in that great silent room. And as long as millions of people continue to access the seemingly endless collection, the Library of Alexandria of the internet will live on, long after its founder, as he puts it, "Goes to the big archive in the sky."

More for you

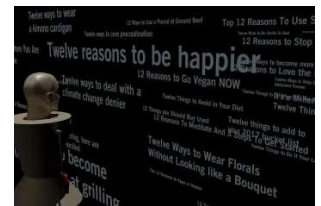
Internet Archive in SF throws a 'Let's fix the fricking web' party

[Read Now](#)



Internet Archive is a treasure trove of material for artists

[Read Now](#)



Chase DiFelicianantonio is a San Francisco Chronicle staff writer. Email: chase.difelicianantonio@sfgate.com Twitter: [@ChaseDiFelice](https://twitter.com/ChaseDiFelice)



Written By
Chase DiFelicianantonio

Reach Chase on



Chase DiFeliciano is a reporter at The San Francisco Chronicle on the Transformation team, where he covers tech culture, workplace safety and labor issues in San Francisco, Silicon Valley and beyond. Prior to joining The Chronicle, he covered immigration for the Daily Journal, a legal affairs newspaper, and a variety of beats at the North Bay Business Journal in Santa Rosa. Chase has degrees in journalism and history from Loyola University Chicago.

[VIEW COMMENTS](#)

Top of the News



Family lost everything in the Caldor Fire. Will they rebuild in the woods?

Generations of Tylers have lived on 500-acres of dense forest in Grizzly Flats in El Dorado County. The Caldor Fire destroyed their homes and heritage when it burned through the town.

BY JULIE JOHNSON

Thank Jimmy G. for being steady, boring — and victorious