THE NEW YORKER

THE COBWEB

Can the Internet be archived?

By Jill Lepore

Malaysia Airlines Flight 17 took off from Amsterdam at 10:31 A.M. G.M.T. on July 17, 2014, for a twelve-hour flight to Kuala Lumpur. Not much more than three hours later, the plane, a Boeing 777, crashed in a field outside Donetsk, Ukraine. All two hundred and ninety-eight people on board were killed. The plane's last radio contact was at 1:20 P.M. G.M.T. At 2:50 P.M. G.M.T., Igor Girkin, a Ukrainian separatist leader also known as Strelkov, or someone acting on his behalf, posted a message on VKontakte, a Russian social-media site: "We just downed a plane, an AN-26." (An Antonov 26 is a Soviet-built military cargo plane.) The post includes links to video of the wreckage of a plane; it appears to be a Boeing 777.

Two weeks before the crash, Anatol Shmelev, the curator of the Russia and Eurasia collection at the Hoover Institution, at Stanford, had submitted to the Internet Archive, a nonprofit library in California, a list of Ukrainian and Russian Web sites and blogs that ought to be recorded as part of the archive's Ukraine Conflict collection. Shmelev is one of about a thousand librarians and archivists around the world who identify possible acquisitions for the Internet Archive's subject collections, which are stored in its Wayback Machine, in San Francisco. Strelkov's VKontakte page was on Shmelev's list. "Strelkov is the field commander in Slaviansk and one of the most important figures in the conflict," Shmelev had written in an e-mail to the Internet Archive on July 1st, and his page "deserves to be recorded twice a day."

On July 17th, at 3:22 P.M. G.M.T., the Wayback Machine saved a screenshot of Strelkov's VKontakte post about downing a plane. Two hours and twenty-two minutes later, Arthur Bright, the Europe editor of the *Christian Science Monitor*, tweeted a picture of the screenshot, along with the message "Grab of Donetsk militant Strelkov's claim of downing what appears to have been MH17." By then, Strelkov's VKontakte

page had already been edited: the claim about shooting down a plane was deleted. The only real evidence of the original claim lies in the Wayback Machine.

The average life of a Web page is about a hundred days. Strelkov's "We just downed a plane" post lasted barely two hours. It might seem, and it often feels, as though stuff on the Web lasts forever, for better and frequently for worse: the embarrassing photograph, the regretted blog (more usually regrettable not in the way the slaughter of civilians is regrettable but in the way that bad hair is regrettable). No one believes any longer, if anyone ever did, that "if it's on the Web it must be true," but a lot of people do believe that if it's on the Web it will stay on the Web. Chances are, though, that it actually won't. In 2006, David Cameron gave a speech in which he said that Google was democratizing the world, because "making more information available to more people" was providing "the power for anyone to hold to account those who in the past might have had a monopoly of power." Seven years later, Britain's Conservative Party scrubbed from its Web site ten years' worth of Tory speeches, including that one. Last year, BuzzFeed deleted more than four thousand of its staff writers' early posts, apparently because, as time passed, they looked stupider and stupider. Social media, public records, junk: in the end, everything goes.

Web pages don't have to be deliberately deleted to disappear. Sites hosted by corporations tend to die with their hosts. When MySpace, GeoCities, and Friendster were reconfigured or sold, millions of accounts vanished. (Some of those companies may have notified users, but Jason Scott, who started an outfit called Archive Team—its motto is "We are going to rescue your shit"—says that such notification is usually purely notional: "They were sending e-mail to dead e-mail addresses, saying, 'Hello, Arthur Dent, your house is going to be crushed.' ") Facebook has been around for only a decade; it won't be around forever. Twitter is a rare case: it has arranged to archive all of its tweets at the Library of Congress. In 2010, after the announcement, Andy Borowitz tweeted, "Library of Congress to acquire entire Twitter archive—will rename itself Museum of Crap." Not long after that, Borowitz abandoned that Twitter account. You might, one day, be able to find his old tweets at the Library of Congress, but not anytime soon: the Twitter Archive is not yet open for research. Meanwhile, on the Web, if you click on a link to Borowitz's tweet about the Museum of Crap, you get this message: "Sorry, that page doesn't exist!"

The Web dwells in a never-ending present. It is—elementally—ethereal, ephemeral, unstable, and unreliable. Sometimes when you try to visit a Web page what you see is an error message: "Page Not Found." This is known as "link rot," and it's a drag, but it's better than the alternative. More often, you see an updated Web page; most likely the original has been overwritten. (To overwrite, in computing, means to destroy old data by storing new data in their place; overwriting is an artifact of an era when computer storage was very expensive.) Or maybe the page has been moved and something else is where it used to be. This is known as "content drift," and it's more pernicious than an error message, because it's impossible to tell that what you're seeing isn't what you went to look for: the overwriting, erasure, or moving of the original is invisible. For the law and for the courts, link rot and content drift, which are collectively known as "reference rot," have been disastrous. In providing evidence, legal scholars, lawyers, and judges often cite Web pages in their footnotes; they expect that evidence to remain where they found it as their proof, the way that evidence on paper—in court records and books and law journals—remains where they found it, in libraries and courthouses. But a 2013 survey of law- and policy-related publications found that, at the end of six years, nearly fifty per cent of the URLs cited in those publications no longer worked. According to a 2014 study conducted at Harvard Law School, "more than 70% of the URLs within the Harvard Law Review and other journals, and 50% of the URLs within United States Supreme Court opinions, do not link to the originally cited information." The overwriting, drifting, and rotting of the Web is no less catastrophic for engineers, scientists, and doctors. Last month, a team of digital library researchers based at Los Alamos National Laboratory reported the results of an exacting study of three and a half million scholarly articles published in science, technology, and medical journals between 1997 and 2012: one in five links provided in the notes suffers from reference rot. It's like trying to stand on quicksand.

The footnote, a landmark in the history of civilization, took centuries to invent and to spread. It has taken mere years nearly to destroy. A footnote used to say, "Here is how I know this and where I found it." A footnote that's a link says, "Here is what I used to know and where I once found it, but chances are it's not there anymore." It doesn't matter whether footnotes are your stock-in-trade. Everybody's in a pinch. Citing a Web page as the source for something you know—using a URL as evidence—is ubiquitous. Many people find themselves doing it three or four times before breakfast and five times more before lunch. What happens when your evidence vanishes by dinnertime?

The day after Strelkov's "We just downed a plane" post was deposited into the Wayback Machine, Samantha Power, the U.S. Ambassador to the United Nations, told the U.N. Security Council, in New York, that Ukrainian separatist leaders had "boasted on social media about shooting down a plane, but later deleted these messages." In San Francisco, the people who run the Wayback Machine posted on the Internet Archive's Facebook page, "Here's why we exist."

The address of the Internet Archive is archive.org, but another way to visit is to take a plane to San Francisco and ride in a cab to the Presidio, past cypresses that look as though someone had drawn them there with a smudgy crayon. At 300 Funston Avenue, climb a set of stone steps and knock on the brass door of a Greek Revival temple. You can't miss it: it's painted wedding-cake white and it's got, out front, eight Corinthian columns and six marble urns.

"We bought it because it matched our logo," Brewster Kahle told me when I met him there, and he wasn't kidding. Kahle is the founder of the Internet Archive and the inventor of the Wayback Machine. The logo of the Internet Archive is a white, pedimented Greek temple. When Kahle started the Internet Archive, in 1996, in his attic, he gave everyone working with him a book called "The Vanished Library," about the burning of the Library of Alexandria. "The idea is to build the Library of Alexandria Two," he told me. (The Hellenism goes further: there's a partial backup of the Internet Archive in Alexandria, Egypt.) Kahle's plan is to one-up the Greeks. The motto of the Internet Archive is "Universal Access to All Knowledge." The Library of Alexandria was open only to the learned; the Internet Archive is open to everyone. In 2009, when the Fourth Church of Christ, Scientist, decided to sell its building, Kahle went to Funston Avenue to see it, and said, "That's our logo!" He loves that the church's cornerstone was laid in 1923: everything published in the United States before that date lies in the public domain. A temple built in copyright's year zero seemed fated. Kahle hops, just slightly, in his shoes when he gets excited. He says, showing me the church, "It's *Greek!*"

Kahle is long-armed and pink-cheeked and public-spirited; his hair is gray and frizzled. He wears round wire-rimmed eyeglasses, linen pants, and patterned button-down shirts. He looks like Mr. Micawber, if Mr. Micawber had left Dickens's London in a time machine and landed in the Pacific, circa 1955, disguised as an American tourist.

Instead, Kahle was born in New Jersey in 1960. When he was a kid, he watched "The Rocky and Bullwinkle Show"; it has a segment called "Peabody's Improbable History," which is where the Wayback Machine got its name. Mr. Peabody, a beagle who is also a Harvard graduate and a Nobel laureate, builds a WABAC machine—it's meant to sound like a UNIVAC, one of the first commercial computers—and he uses it to take a boy named Sherman on adventures in time. "We just set it, turn it on, open the door, and there we are—or *were*, really," Peabody says.

When Kahle was growing up, some of the very same people who were building what would one day become the Internet were thinking about libraries. In 1961, in Cambridge, J. C. R. Licklider, a scientist at the technology firm Bolt, Beranek and Newman, began a two-year study on the future of the library, funded by the Ford Foundation and aided by a team of researchers that included Marvin Minsky, at M.I.T. As Licklider saw it, books were good at displaying information but bad at storing, organizing, and retrieving it. "We should be prepared to reject the schema of the physical book itself," he argued, and to reject "the printed page as a long-term storage device." The goal of the project was to imagine what libraries would be like in the year 2000. Licklider envisioned a library in which computers would replace books and form a "network in which every element of the fund of knowledge is connected to every other element."

In 1963, Licklider became a director at the Department of Defense's Advanced Research Projects Agency (now called DARPA). During his first year, he wrote a seven-page memo in which he addressed his colleagues as "Members and Affiliates of the Intergalactic Computer Network," and proposed the networking of ARPA machines. This sparked the imagination of an electrical engineer named Lawrence Roberts, who later went to ARPA from M.I.T.'s Lincoln Laboratory. (Licklider had helped found both B.B.N. and Lincoln.) Licklider's two-hundred-page Ford Foundation report, "Libraries of the Future," was published in 1965. By then, the network he imagined was already being built, and the word "hyper-text" was being used. By 1969, relying on a data-transmission technology called "packet-switching" which had been developed by a Welsh scientist named Donald Davies, ARPA had built a computer network called ARPANET. By the mid-nineteen-seventies, researchers across the country had developed a network of networks: an internetwork, or, later, an "internet."

Kahle enrolled at M.I.T. in 1978. He studied computer science and engineering with Minsky. After graduating, in 1982, he worked for and started companies that were later sold for a great deal of money. In the late eighties, while working at Thinking Machines, he developed Wide Area Information Servers, or WAIS, a protocol for searching, navigating, and publishing on the Internet. One feature of WAIS was a time axis; it provided for archiving through version control. (Wikipedia has version control; from any page, you can click on a tab that says "View History" to see all earlier versions of that page.) WAIS came before the Web, and was then overtaken by it. In 1989, at CERN, the European Particle Physics Laboratory, in Geneva, Tim Berners-Lee, an English computer scientist, proposed a hypertext transfer protocol (HTTP) to link pages on what he called the World Wide Web. Berners-Lee toyed with the idea of a time axis for his protocol, too. One reason it was never developed was the preference for the most up-to-date information: a bias against obsolescence. But the chief reason was the premium placed on ease of use. "We were so young then, and the Web was so young," Berners-Lee told me. "I was trying to get it to go. Preservation was not a priority. But we're getting older now." Other scientists involved in building the infrastructure of the Internet are getting older and more concerned, too. Vint Cerf, who worked on ARPANET in the seventies, and now holds the title of Chief Internet Evangelist at Google, has started talking about what he sees as a need for "digital vellum": long-term storage. "I worry that the twenty-first century will become an informational black hole," Cerf e-mailed me. But Kahle has been worried about this problem all along.

"I'm completely in praise of what Tim Berners-Lee did," Kahle told me, "but he kept it very, very simple." The first Web page in the United States was created at SLAC, Stanford's linear-accelerator center, at the end of 1991. Berners-Lee's protocol—which is not only usable but also elegant—spread fast, initially across universities and then into the public. "Emphasized text like *this* is a hypertext link," a 1994 version of SLAC's Web page explained. In 1991, a ban on commercial traffic on the Internet was lifted. Then came Web browsers and e-commerce: both Netscape and Amazon were founded in 1994. The Internet as most people now know it—Web-based and commercial—began in the mid-nineties. Just as soon as it began, it started disappearing.

And the Internet Archive began collecting it. The Wayback Machine is a Web archive, a collection of old Web pages; it is, in fact, *the* Web archive. There are

others, but the Wayback Machine is so much bigger than all of them that it's very nearly true that if it's not in the Wayback Machine it doesn't exist. The Wayback Machine is a robot. It crawls across the Internet, in the manner of Eric Carle's very hungry caterpillar, attempting to make a copy of every Web page it can find every two months, though that rate varies. (It first crawled over this magazine's home page, newyorker.com, in November, 1998, and since then has crawled the site nearly seven thousand times, lately at a rate of about six times a day.) The Internet Archive is also stocked with Web pages that are chosen by librarians, specialists like Anatol Shmelev, collecting in subject areas, through a service called Archive It, at archive-it.org, which also allows individuals and institutions to build their own archives. (A copy of everything they save goes into the Wayback Machine, too.) And anyone who wants to can preserve a Web page, at any time, by going to archive.org/web, typing in a URL, and clicking "Save Page Now." (That's how most of the twelve screenshots of Strelkov's VKontakte page entered the Wayback Machine on the day the Malaysia Airlines flight was downed: seven captures that day were made by a robot; the rest were made by humans.)

I was on a panel with Kahle a few years ago, discussing the relationship between material and digital archives. When I met him, I was struck by a story he told about how he once put the entire World Wide Web into a shipping container. He just wanted to see if it would fit. How big is the Web? It turns out, he said, that it's twenty feet by eight feet by eight feet, or, at least, it was on the day he measured it. How much did it weigh? Twenty-six thousand pounds. He thought that *meant* something. He thought people needed to *know* that.

Kahle put the Web into a storage container, but most people measure digital data in bytes. This essay is about two hundred thousand bytes. A book is about a megabyte. A megabyte is a million bytes. A gigabyte is a billion bytes. A terabyte is a million million bytes. A petabyte is a million gigabytes. In the lobby of the Internet Archive, you can get a free bumper sticker that says "10,000,000,000,000,000 Bytes Archived." Ten petabytes. It's obsolete. That figure is from 2012. Since then, it's doubled.

The Wayback Machine has archived more than four hundred and thirty billion Web pages. The Web is global, but, aside from the Internet Archive, a handful of fledgling commercial enterprises, and a growing number of university Web archives, most Web

archives are run by national libraries. They collect chiefly what's in their own domains (the Web Archive of the National Library of Sweden, for instance, includes every Web page that ends in ".se"). The Library of Congress has archived nine billion pages, the British Library six billion. Those collections, like the collections of most national libraries, are in one way or another dependent on the Wayback Machine; the majority also use Heritrix, the Internet Archive's open-source code. The British Library and the Bibliothèque Nationale de France backfilled the early years of their collections by using the Internet Archive's crawls of the .uk and .fr domains. The Library of Congress doesn't actually do its own Web crawling; it contracts with the Internet Archive to do it instead.

The church at 300 Funston Avenue is twenty thousand square feet. The Internet Archive, the building, is open to the public most afternoons. It is, after all, a library. In addition to housing the Wayback Machine, the Internet Archive is a digital library, a vast collection of digitized books, films, television and radio programs, music, and other stuff. Because of copyright, not everything the Internet Archive has digitized is online. In the lobby of the church, there's a scanning station and a listening room: two armchairs, a coffee table, a pair of bookshelves, two iPads, and two sets of headphones. "You can listen to anything here," Kahle says. "We can't put all our music on the Internet, but we can put everything here."

Copyright is the elephant in the archive. One reason the Library of Congress has a very small Web-page collection, compared with the Internet Archive, is that the Library of Congress generally does not collect a Web page without asking, or, at least, giving notice. "The Internet Archive hoovers," Abbie Grotke, who runs the Library of Congress's Web-archive team, says. "We can't hoover, because we have to notify site owners and get permissions." (There are some exceptions.) The Library of Congress has something like an opt-in policy; the Internet Archive has an opt-out policy. The Wayback Machine collects every Web page it can find, unless that page is blocked; blocking a Web crawler requires adding only a simple text file, "robots.txt," to the root of a Web site. The Wayback Machine will honor that file and not crawl that site, and it will also, when it comes across a robots.txt, remove all past versions of that site. When the Conservative Party in Britain deleted ten years' worth of speeches from its Web site, it also added a robots.txt, which meant that, the next time the Wayback Machine tried to crawl the site, all its captures of those speeches went away, too. (Some have since

been restored.) In a story that ran in the *Guardian*, a Labour Party M.P. said, "It will take more than David Cameron pressing delete to make people forget about his broken promises." And it would take more than a robots.txt to entirely destroy those speeches: they have also been collected in the U.K. Web Archive, at the British Library. The U.K. has what's known as a legal-deposit law; it requires copies of everything published in Britain to be deposited in the British Library. In 2013, that law was revised to include everything published on the U.K. Web. "People put their private lives up there, and we actually don't want that stuff," Andy Jackson, the technical head of the U.K. Web Archive, told me. "We don't want anything that you wouldn't consider a publication." It is hard to say quite where the line lies. But Britain's legal-deposit laws mean that the British Library doesn't have to honor a request to stop collecting.

MORE FROM THIS ISSUE

JANUARY 26, 2015

CULTURAL CHRONICLES

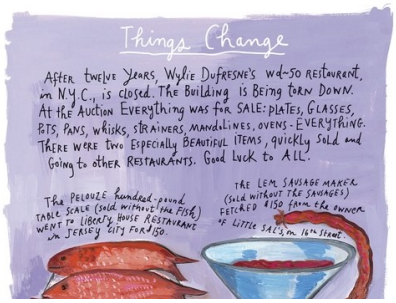
The Next Thing

By Adam Gopnik

SHOUTS & MURMURS

To Fall Out of Love, Do This

By Susanna Wolff



SKETCHBOOK

Things Change

By Maira Kalman

BOOKS

Briefly

Legal-deposit laws have been the standard in Western Europe for centuries. They provide national libraries with a form of legal protection unavailable to the Library of Congress, which is not strictly a national library; also, U.S. legal- deposit laws have exempted online-only works. "We are citadels," Gildas Illien, the former Web archivist at the Bibliothèque Nationale de France, told me. The Internet Archive is an invaluable public institution, but it's not a national library, either, and, because the law of copyright

has not kept up with technological change, Kahle has been collecting Web sites and making them freely available to the public without the full and explicit protection of the law. "It's extremely audacious," Illien says. "In Europe, no organization, or very few, would take that risk." There's another feature to legal-deposit laws like those in France, a compromise between advocates of archiving and advocates of privacy. Archivists at the BnF can capture whatever Web pages they want, but those collections can be used only in the physical building itself. (For the same reason, you can't check a book out of the Bibliothèque Nationale de France; you have to read it there.) One result is that the BnF's Web archive is used by a handful of researchers, a few dozen a month; the Wayback Machine is used by hundreds of thousands of people a day.

In 2002, Kahle proposed an initiative in which the Internet Archive, in collaboration with national libraries, would become the head of a worldwide consortium of Web archives. (The Internet Archive collects from around the world, and is available in most of the world. Currently, the biggest exception is China—"I guess because we have materials on the archive that the Chinese government would rather not have its citizens see," Kahle says.) This plan didn't work out, but from that failure came the International Internet Preservation Consortium, founded in 2003 and chartered at the BnF. It started with a dozen member institutions; there are now forty-nine.

Something else came out of that consortium. I talked to Illien two days after the massacre in Paris at the offices of *Charlie Hebdo*. "We are overwhelmed, and scared, and even taking the subway is terrifying, and we are scared for our children," Illien said. "The library is a target." When we spoke, the suspects were still at large; hostages had been taken. Illien and his colleagues had started a Web archive about the massacre and the world's response. "Right now the media is full of it, but we know that most of that won't last," he said. "We wrote to our colleagues around the world and asked them to send us feeds to these URLs, to Web sites that were happening, right now, in Paris, so that we could collect them and historians will one day be able to see." He was very quiet. He said, "When something like that happens, you wonder what you can do from where you sit. Our job is memory."

The plan to found a global Internet archive proved unworkable, partly because national laws relating to legal deposit, copyright, and privacy are impossible to reconcile, but also because Europeans tend to be suspicious of American organizations based in Silicon

Valley ingesting their cultural inheritance. Illien told me that, when faced with Kahle's proposal, "national libraries decided they could not rely on a third party," even a nonprofit, "for such a fundamental heritage and preservation mission." In this same spirit, and in response to Google Books, European libraries and museums collaborated to launch Europeana, a digital library, in 2008. The Googleplex, Google's headquarters, is thirty-eight miles away from the Internet Archive, but the two could hardly be more different. In 2009, after the Authors Guild and the Association of American Publishers sued Google Books for copyright infringement, Kahle opposed the proposed settlement, charging Google with effectively attempting to privatize the public-library system. In 2010, he was on the founding steering committee of the Digital Public Library of America, which is something of an American version of Europeana; its mission is to make what's in libraries, archives, and museums "freely available to the world . . . in the face of increasingly restrictive digital options."

Kahle is a digital utopian attempting to stave off a digital dystopia. He views the Web as a giant library, and doesn't think it ought to belong to a corporation, or that anyone should have to go through a portal owned by a corporation in order to read it. "We are building a library that is us," he says, "and it is ours."

When the Internet Archive bought the church, Kahle recalls, "we had the idea that we'd convert it into a library, but what does a library look like anymore? So we've been settling in, and figuring that out."

From the lobby, we headed up a flight of yellow-carpeted stairs to the chapel, an enormous dome-ceilinged room filled with rows of oak pews. There are arched stained-glass windows, and the dome is a stained-glass window, too, open to the sky, like an eye of God. The chapel seats seven hundred people. The floor is sloped. "At first, we thought we'd flatten the floor and pull up the pews," Kahle said, as he gestured around the room. "But we couldn't. They're just too beautiful."

On the wall on either side of the altar, wooden slates display what, when this was a church, had been the listing of the day's hymn numbers. The archivists of the Internet have changed those numbers. One hymn number was 314. "Do you know what that is?" Kahle asked. It was a test, and something of a trick question, like when someone asks you what's your favorite B track on the White Album. "Pi," I said, dutifully, or its first three digits, anyway. Another number was 42. Kahle gave me an inquiring look. I rolled

my eyes. Seriously? But it is serious, in a way. It's hard not to worry that the Wayback Machine will end up like the computer in Douglas Adams's "Hitchhiker's Guide to the Galaxy," which is asked what is the meaning of "life, the universe, and everything," and, after thinking for millions of years, says, "Forty-two." If the Internet can be archived, will it ever have anything to tell us? Honestly, isn't most of the Web trash? And, if everything's saved, won't there be too much of it for anyone to make sense of any of it? Won't it be useless?

The Wayback Machine is humongous, and getting humongouser. You can't search it the way you can search the Web, because it's too big and what's in there isn't sorted, or indexed, or catalogued in any of the many ways in which a paper archive is organized; it's not ordered in any way at all, except by URL and by date. To use it, all you can do is type in a URL, and choose the date for it that you'd like to look at. It's more like a phone book than like an archive. Also, it's riddled with errors. One kind is created when the dead Web grabs content from the live Web, sometimes because Web archives often crawl different parts of the same page at different times: text in one year, photographs in another. In October, 2012, if you asked the Wayback Machine to show you what cnn.com looked like on September 3, 2008, it would have shown you a page featuring stories about the 2008 McCain-Obama Presidential race, but the advertisement alongside it would have been for the 2012 Romney-Obama debate. Another problem is that there is no equivalent to what, in a physical archive, is a perfect provenance. Last July, when the computer scientist Michael Nelson tweeted the archived screenshots of Strelkov's page, a man in St. Petersburg tweeted back, "Yep. Perfect tool to produce 'evidence' of any kind." Kahle is careful on this point. When asked to authenticate a screenshot, he says, "We can say, 'This is what we know. This is what our records say. This is how we received this information, from which apparent Web site, at this IP address.' But to actually say that this happened in the past is something that we can't say, in an ontological way." Nevertheless, screenshots from Web archives have held up in court, repeatedly. And, as Kahle points out, "They turn out to be much more trustworthy than most of what people try to base court decisions on."

You can do something more like keyword searching in smaller subject collections, but nothing like Google searching (there is no relevance ranking, for instance), because the tools for doing anything meaningful with Web archives are years behind the tools for creating those archives. Doing research in a paper archive is to doing research in a Web

archive as going to a fish market is to being thrown in the middle of an ocean; the only thing they have in common is that both involve fish.

The Web archivists at the British Library had the brilliant idea of bringing in a team of historians to see what they could do with the U.K. Web Archive; it wasn't all that much, but it was helpful to see what they *tried* to do, and why it didn't work. Gareth Millward, a young scholar interested in the history of disability, wanted to trace the history of the Royal National Institute for the Blind. It turned out that the institute had endorsed a talking watch, and its name appeared in every advertisement for the watch. "This one advert appears thousands of times in the database," Millward told me. It cluttered and bogged down nearly everything he attempted. Last year, the Internet Archive made an archive of its .gov domain, tidied up and compressed the data, and made it available to a group of scholars, who tried very hard to make something of the material. It was so difficult to recruit scholars to use the data that the project was mostly a wash. Kahle says, "I give it a B." Stanford's Web archivist, Nicholas Taylor, thinks it's a chicken-and-egg problem. "We don't know what tools to build, because no research has been done, but the research hasn't been done because we haven't built any tools."

The footnote problem, though, stands a good chance of being fixed. Last year, a tool called Perma.cc was launched. It was developed by the Harvard Library Innovation Lab, and its founding supporters included more than sixty law-school libraries, along with the Harvard Berkman Center for Internet and Society, the Internet Archive, the Legal Information Preservation Alliance, and the Digital Public Library of America. Perma.cc promises "to create citation links that will never break." It works something like the Wayback Machine's "Save Page Now." If you're writing a scholarly paper and want to use a link in your footnotes, you can create an archived version of the page you're linking to, a "permalink," and anyone later reading your footnotes will, when clicking on that link, be brought to the permanently archived version. Perma.cc has already been adopted by law reviews and state courts; it's only a matter of time before it's universally adopted as the standard in legal, scientific, and scholarly citation.

Perma.cc is a patch, an excellent patch. Herbert Van de Sompel, a Belgian computer scientist who works at the Los Alamos National Laboratory, is trying to reweave the fabric of the Web. It's not possible to go back in time and rewrite the HTTP protocol, but Van de Sompel's work involves adding to it. He and Michael Nelson are part of the

team behind Memento, a protocol that you can use on Google Chrome as a Web extension, so that you can navigate from site to site, and from time to time. He told me, "Memento allows you to say, 'I don't want to see this link where it points me to today; I want to see it around the time that this page was written, for example.' " It searches not only the Wayback Machine but also every major public Web archive in the world, to find the page closest in time to the time you'd like to travel to. ("A world with one archive is a really bad idea," Van de Sompel points out. "You need redundance.") This month, the Memento group is launching a Web portal called Time Travel. Eventually, if Memento and projects like it work, the Web will have a time dimension, a way to get from now to then, effortlessly, a fourth dimension. And then the past will be inescapable, which is as terrifying as it is interesting.

At the back of the chapel, up a short flight of stairs, there are two niches, arched alcoves the same shape and size as the stained-glass windows. Three towers of computers stand within each niche, and ten computers are stacked in each tower: black, rectangular, and humming. There are towers like this all over the building; these are only six of them. Still, this is *it*.

Kahle stands on his tiptoes, sinks back into his sneakers, and then bounds up the stairs. He is like a very sweet boy who, having built a very fine snowman, drags his mother outdoors to see it before it melts. I almost expect him to take my hand. I follow him up the stairs.

"Think of them as open stacks," he says, showing me the racks. "You can walk right up to them and touch them." He reaches out and traces the edge of one of the racks with the tip of his index finger. "If you had all the words in every book in the Library of Congress, it would be about an inch, here," he says, measuring the distance between his forefinger and thumb.

Up close, they're noisy. It's mainly fans, cooling the machines. At first, the noise was a problem: a library is supposed to be quiet. Kahle had soundproofing built into the walls.

Each unit has a yellow and a green light, glowing steadily: power indicators. Then, there are blue lights, flickering.

"Every time a light blinks, someone is uploading or downloading," Kahle explains. Six hundred thousand people use the Wayback Machine every day, conducting two thousand searches a second. "You can *see* it." He smiles as he watches. "They're glowing books!" He waves his arms. "They glow when they're being read!"

One day last summer, a missile was launched into the sky and a plane crashed in a field. "We just downed a plane," a soldier told the world. People fell to the earth, their last passage. Somewhere, someone hit "Save Page Now."

Where is the Internet's memory, the history of our time?

"It's right *here!*" Kahle cries.

The machine hums and is muffled. It is sacred and profane. It is eradicable and unbearable. And it glows, against the dark. ♦

Jill Lepore is a staff writer at The New Yorker and a professor of history at Harvard University. Her latest book, "These Truths: A History of the United States," came out in September. *Read more »*

Read something that means something. Join *The New Yorker* and get a free tote. Subscribe now. »

CONDÉ NAST

© 2019 Condé Nast. All rights reserved. Use of and/or registration on any portion of this site constitutes acceptance of our User Agreement (updated 5/25/18) and Privacy Policy and Cookie Statement (updated 5/25/18). Your California Privacy Rights. The material on this site may not be reproduced, distributed, transmitted, cached or otherwise used, except with the prior written permission of Condé Nast. *The New Yorker* may earn a portion of sales from products and services that are purchased through links on our site as part of our affiliate partnerships with retailers. Ad Choices