

ANNALS OF TECHNOLOGY JANUARY 26, 2015 ISSUE

THE COBWEB

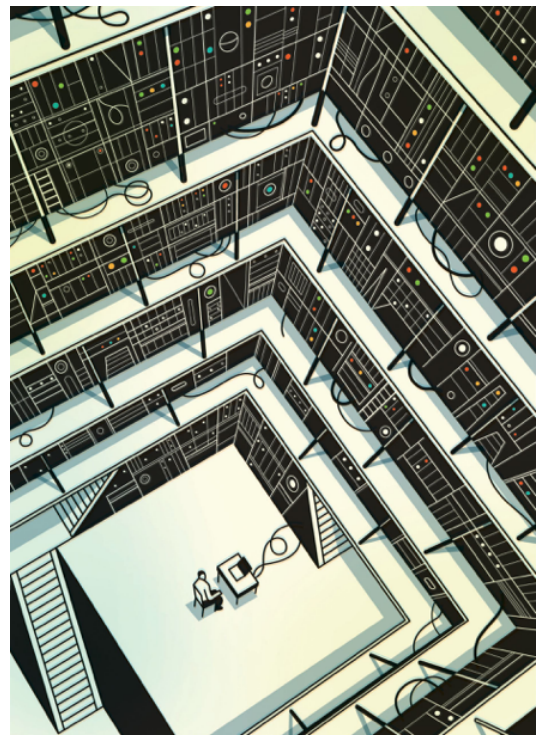
Can the Internet be archived?**By Jill Lepore**

January 19, 2015

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



The Web wasn't built to preserve its past; the Wayback Machine aims to remedy that. Illustration by Harry Campbell

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](http://www.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 hovers,” Abbie Grotke, who runs the Library of Congress’s Web-archive team, says. “We can’t hover, 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.

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. ♦

Published in the print edition of the January 26, 2015, issue.



Jill Lepore is a professor of history at Harvard and the host of the podcast “The Last Archive.” Her fourteenth book, “If Then,” will be published in September.

The newsmagazine of the American Historical Association

PERSPECTIVES ON HISTORY

VISIT AHA



[About](#)

[Research](#)

[Teaching & Learning](#)

[Professional Life](#)

[All Topics](#)

[Magazine](#)

[Home](#) > [Publications & Directories](#) > [Perspectives on History](#)
> [Issues](#) > [November 2016](#) > News > Doing Right Online:
Archivists Shape an Ethics for the Digital Age

NEWS

DOING RIGHT ONLINE: ARCHIVISTS SHAPE AN ETHICS FOR THE DIGITAL AGE

Kritika Agarwal | Nov 1, 2016



IN THIS SECTION

About

Research

Teaching & Learning

Professional Life

All Topics

Magazine



Black Lives Matter protesters at the Minnesota Governor's Mansion in July 2016. Archivists at DocNow work with community activists to document the offline labor that makes social media hashtag campaigns such as #BlackLivesMatter possible. Tony Webster/Flickr/CC BY-SA 2.0

At a time when the CIA invests in companies that develop surveillance technologies for social media, archivists like Bergis Jules face disconcerting challenges. An archivist at the University of California, Riverside, Jules is also community lead on Documenting the Now (DocNow). This digital project brings together archivists, academics, and activists to create ethical standards for the archiving of tweets related to the Black Lives Matter and other social justice movements, so the matter of surveillance is not just a theoretical concern. For those involved in DocNow, the possibility that their archival efforts will be used in police surveillance is an ethical matter they must confront. Archivists, Jules says, must actively think about “how . . . the collections with social media content that we build might support law enforcement activity that targets groups of people they don’t agree with—for example, activists.”

Surveillance is only one concern of archivists who build digital collections. The availability of digital records has proved a boon for historians (for example, by reducing

costs and overcoming distance), but for archivists, the ease of access that digitization brings also provokes a host of ethical concerns about what to digitize and how to do it. Some questions, such as those of gaining consent from content creators before displaying materials online or ensuring that materials are presented in their appropriate context in the digital realm, are reiterations of old problems. Others, such as those of online surveillance and digital privacy, are very much the products of the 21st century. As archivists forge practices for ethical online behavior, some are discovering new uses of digital technology that can rectify injustices associated with historic collection and archiving practices.

Michelle Caswell, who teaches archival theory at UCLA and cofounded the South Asian American Digital Archive, advises archivists to consider “whether the record creators and subjects of those records would consent to having them available digitally.” While archivists typically seek consent to make materials publicly available for historical research, what makes the issue thornier in the case of digital collections is the expansion of the meaning of “the public.” According to Caswell, it is one thing for a record to be available publicly in a repository, at which a researcher has to physically show up and request materials, and another for it to be searchable and discoverable by anyone in the world with an Internet connection. As Jules points out, creating a digital archive essentially creates a collection of digital data, which researchers can mine in ways that go well beyond what is possible with physical collections.

To take one example, Reveal Digital, a website that uses a crowd-funded model to digitize archival collections, attracted criticism recently for digitizing back issues of the historic feminist lesbian porn magazine *On Our Backs*, held in special collections at Duke and Northwestern Universities. According to Tara Robertson, a systems librarian and accessibility advocate, even though Reveal Digital claims to have obtained permission

from relevant copyright holders, it did not seek consent from the individual contributors to the magazine. The very act of digitization, according to Robertson, placed at risk subjects of porn shoots who had probably never envisioned the magazine to be so publicly available and searchable. One of the subjects who appeared in the magazine told Robertson, “When I heard all the issues of the magazine are being digitized, my heart sank. I meant this work to be for my community, and now I am being objectified in a way that I have no control over.” Another subject, who appeared on the cover of the magazine, worried that having the content freely available online would impact her professional career in the technology industry.

To address these issues, some archivists seek to explicitly gather consent from content creators before placing it online, and in doing so, they go above and beyond what is required of them under copyright law. In other instances, explains Cathy Moran Hajo, director of the Jane Addams Papers Project at Ramapo College, material might be posted online but with redacted personal information. Take-down policies also allow users to request removal of objectionable materials. As a demonstration of how seriously it takes matters of consent, DocNow is working to create a system that would allow Twitter users to opt out of having their tweets archived, though they are publicly available.

Another ethical concern that goes hand in hand with consent is that of context—ensuring that digital materials presented online are not isolated from the circumstances in which they were created. For DocNow, that means recognizing and documenting the offline activism that made social media hashtag campaigns such as #BlackLivesMatter possible. In order to do this, DocNow is actively engaging with community activists to learn how they want their online activism to be remembered and archived.

Instead of using an existing digital archival system and then working within its constraints, DocNow is letting ethical concerns drive its creation of technology. It isn't alone. An increasing number of archivists and scholars are now using digital tools and technology to confront ethical issues that have historically plagued collection and archiving practices. At the forefront of these efforts are archivists working with indigenous peoples and collections. As Kim Christen Withey (Washington State Univ.) put it in a recent panel discussion at the Library of Congress, "The history of collection is the history of colonialism." Indigenous peoples rarely hold copyright to materials related to their cultural or ancestral heritage held at libraries and archives around the world, and as Caswell explains, many of these records "were created without the consent of the indigenous communities" and "contain sacred information that was never meant to be distributed on a wider basis." In response, many libraries, archives, and museums are not only rethinking the widely accepted ethos of "open access" in the archival world; they are also moving to a collaborative approach, working with indigenous communities to obtain permissions and to gather contextual information or create metadata.

One of the most forward-thinking and innovative of these collaborative approaches is an online content management system named Mukurtu. Managed by the Center for Digital Scholarship and Curation at Washington State University and directed by Withey, Mukurtu offers a platform that allows indigenous communities to digitally archive their heritage and knowledge, granting access to some users while restricting it to others. For example, using Mukurtu, an indigenous community can determine whether an image of a sacred object should be available publicly or only to a few registered users. An extension of Mukurtu is the Traditional Knowledge (TK) labels tool, which allows universities and libraries to add labels to digital materials to add context and indicate an indigenous community's

preference for how researchers should view and use cultural materials. The Library of Congress plans to use the TK labels as part of its forthcoming digital collection of original wax cylinder recordings from the Passamaquoddy people made in 1890 by anthropologist Jesse Walter Fewkes.

For historians, consideration of ethical concerns surrounding consent, context, and access could mean shouldering some of the responsibility of ensuring ethical use of archival materials, whether traditional or born-digital. Researchers might need to weigh whether a particular archival material is ethical for them to use, keeping in mind that most research now ends up online. Hajo recalls that archivists working on the Margaret Sanger Papers Project redacted the name of a woman who received an abortion in a birth control clinic from microfilmed records, but a scholar using the physical papers published the woman's name, causing it to appear on Google Books. Thus, even when something is publicly available, like a tweet, scholars might need to make ethical choices about using and presenting that information. Philippa Levine, vice president of the AHA's Professional Division, says that historians should indeed consider these questions as they navigate the new avenues of research opened up by digital holdings. Jules also encourages historians to get involved in the process of creating digital collections and in discussions of ethical concerns. "Be part of the conversation," he says.

The stakes are undoubtedly high. In September 2016, the *Baltimore Sun* reported that its police force had used the service Geofeedia, which analyzes social media information "to monitor protests, parades, and holiday celebrations." In October, the American Civil Liberties Union released a report noting that the use of such software was more widespread than previously thought. Ensuring that archivists and historians do not become complicit in the marginalization of vulnerable

populations because of their online practices is certainly an ethical conversation worth having.

Kritika Agarwal is associate editor, publications, at the AHA. She tweets @kritikaldesi.



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](#). Attribution must provide author name, article title, *Perspectives on History*, date of publication, and a link to this page. This license applies only to the article, not to text or images used here by permission.

The American Historical Association welcomes comments in the discussion area below, at [AHA Communities](#), and in [letters to the editor](#). Please read our [commenting and letters policy](#) before submitting.

Tags:[History News](#)[Archives](#)[Digital History](#)[Thematic](#)[Indigenous History](#)[Public History](#)

COMMENT

Please read our [commenting and letters policy](#) before submitting.

0 Comments Sort by Oldest

Add a comment...

[Facebook Comments Plugin](#)

AHA SITE MAP

- About
- Research
- Teaching & Learning
- Professional Life
- All Topics
- Magazine
- Full Site Map**

GET INVOLVED

Why should I join the AHA?

How can I support the AHA?



CONTACT

Phone: 202.544.2422
Email: info@historians.org

**400 A Street SE
Washington, DC 20003**

© 2018 American
Historical Association

important. Since the eighteenth century, restrictive archives both in the United States and in Europe have modified their procedures in the name of democratic openness. The tension between openness and privilege appeared to be in equilibrium. Then the information technology revolution led to an entirely new set of issues that soon pervaded the archival and information business. The large-scale digital library brings with it another set of access dilemmas. And information technology is combining with several other seemingly unstoppable forces—and transforming the ethics of archives. Mapping hard-won and long-cherished values over into the digital environment is the new challenge.

Proprietary Control vs. Free Access in the Digital Environment

For half a millenium, since the Gutenberg revolution in the fifteenth century, there has been a clear distinction between a published book and a manuscript. The book allows multiple identical copies to be distributed among many readers in many locations. Manuscripts since antiquity have been less stable objects: unique originals have been laboriously copied with many alterations and variants. The manuscripts that serve as the basis of books typically undergo numerous drafts and changes prior to freezing into the more fixed print version. Where do cumulative online texts fit into this schema? Take for example the articles in Wikipedia that are subject to constant revision. It is no accident that information technology (IT) professionals call these unstable texts *documents* and have turned the word *archive* into a verb describing ways to save text in all its variants. This digital format shares many attributes with archives and is increasingly a subject for the archival profession. Is digital text a publication or a document?

American courts have interpreted digitized documents as publications. There is logic to this viewpoint. In the case of the cigarette papers, this interpretation of digitized text as a publication, one that is protected by the First Amendment, saved the archives from surrendering pirated copies. Microfilm versions of archives have long been called *publications*.

Both online and microform versions of archives allow the text to be distributed among many readers in many locations—the very definition of a publication.

Digital texts, unlike books and microfilm, are subject to many revisions and changes, even tampering and hostile vandalism. Migration from one platform to another can introduce both intended and unintended alterations in format and text. These digital objects typically exist in many versions with deletions and additions—much like a manuscript.

To establish an authoritative version of a literary masterpiece, an editor must determine the correct chronological sequence of manuscript drafts, and then study related documentation to determine the author's final intent. It is recognized as a complex editorial job to establish a "critical edition." With digital texts, it can also be a challenge to find, analyze, and properly sequence the variants. In a way, digital texts, because of their extraordinary mutability and wide distribution, are even more challenging. One can even conceptualize digital texts as "permanently variable" manuscript drafts. Certainly the articles in Wikipedia are intended to be constantly in flux, constantly debated, corrected, and expanded over time. With Wikipedia, these changes are consciously tracked in a detailed and transparent way as the text is constantly corrected and theoretically improved.³⁵ (It should be noted that erroneous information extracted from these texts can be embedded in other documents that remain uncorrected as the source article is updated.) For most web pages, tracing earlier versions is a hit and miss operation, depending on such things as the "Wayback Machine" of the San Francisco-based Internet Archives and cached pages on Google. These efforts at fixing ephemeral and mutating texts are remarkable for what they can do, but in the end they cannot capture deep web structures, password-protected sites, or subscription based web pages.

The version of text you see today may not be the one your colleague told you about yesterday. Authentication is again a major preoccupation, as it was in the pre-Gutenberg era. As online text increasingly supplements and replaces paper-based text, even in the newspaper field, digital objects will have more in common with manuscripts and archives than with print publications. Abby Smith reached a similar conclusion. She observed that

many digital sites are “collections of archival materials that, in the analog realm, would go to a special collections library without being published.”³⁶ It is interesting to see how much of traditional archival practice translates over to digital formats. It is also interesting to see how many longstanding archival ethical concerns also have pertinence in the digital realm. Authenticity is a major one, as are issues concerning privacy, piracy, and—especially—equal access. In fact, in each of these areas, including access, the digital format seems to amplify the scope of the traditional analog problem.

The question needs to be asked: should archivists concern themselves with databases and digital objects? Once archival material has been scanned and entered into this format, it is now a different genre in worldwide distribution, no longer contained by the traditional archives reading room. Should these tools be left in the hands of IT experts who know so much about the highly convoluted coding that is involved? The question is answering itself. Archivists have readily accepted computerized versions of finding aids, and then watched as these tools are filled out with full text online, as is happening with the evolving Online Archive of California. Archivists are required to facilitate the reformatting of materials to digital form and help with the search functions. Archivists must be prepared to appraise born-digital documentation in the originating offices, and confront it increasingly in new acquisitions. In these roles, archivists are already participating as advocates for open information in a rapidly evolving digital world.

The answer to the question, then, is *yes*; archivists need to be actively involved in the management of digital primary sources. They need to advise on the architecture of databases to preserve the two simple principles of open and equal access. Both the technical and the commercial sides of this complex of issues need to be confronted directly. There is good news and bad news on this front. The story starts with the latter, but it ends with some very promising developments.

With this digital revolution, the ground under archivists is shifting. It is not entirely clear how things will unfold, but very powerful societal forces are at work. They are irrevocably changing the information environment.

The forces behind this transformation are linked to the rapid development of computer technology since the introduction of the World Wide Web in 1993–1994. The networked global environment is impacting the flow of information—the way information is recorded, stored, accessed, and restricted. There are both financial and political consequences. The globalized marketplace impacts archival practice. This effect is easy to visualize. The growth of commercial packaging of archives has led to the commodification of the archives as products for licensing as commercial databases. Another effect that is basically political and more difficult to manage is the way real and perceived security threats impact the flow of information. Unpredictable and unfamiliar forms of terrorism and globalized gangsterism have emerged in the twenty-first century. Throughout history, security threats and political challenges have created the dual danger of intrusive monitoring of private information and aggressive censorship of public information: in effect, controlling access to information. In response to the 9/11 tragedy, the U.S. government decided it needed to monitor communications more extensively and intrusively. In response to political challenges, China has experimented with limiting access to certain sites on the Internet in order to squelch opposition. These factors are all interconnected with the advances in information technology. As information brokers, archivists are at the center of the IT storm.

The digital environment impacts responsible access in unpredictable ways. There is more information available, but less reliability and more efforts at controlling it. These forces are likely to continue impacting access, even as a huge influx of digital documentation is transferred to archives. The fragility of digital archives is a pressing concern. The ultimate restriction on access is the permanent loss of documentation. Digital preservation is a highly technical subject, certainly beyond the scope of a book on ethics. What is relevant to archival ethics is the urgent need to protect cultural heritage that is in digital form from major threats: technical, commercial, and political. Digital data is heavily mediated by a technical interface. To be read that interface needs to be maintained, and maintenance is expensive. One can visualize the issue as the need to repurchase ballooning digital

archives every four or five years. Both the cost and the volume of digital documents are threats to access.

New archival acquisitions are arriving in repositories as electronic databases—each with its own quirks, each with masses of unsorted data. Some times these electronic formats are scans of data gleaned from paper sources, other times they are born digital. Such databases contain too much material to simply print out, and the printouts would not preserve the interaction function for conducting searches with combinations of keywords. The shift from paper reference to digital is unavoidable. And it is unavoidably complex, which complicates the implementation of the seemingly simple imperative for open and equal access.

The computer industry is driven by the profit motive, which created a tsunami of commercial innovations. Archives are valuable assets. It was only a matter of time for companies that specialize in marketing microform sets and online journals to begin absorbing out-of-print books, and then archival resources.

Commodification of Information and the Ethics of Access

Database technology is presenting a dilemma for advocates of open and equal access during the transition from paper-based to online research modes. There are unprecedented opportunities for open access to data never before accessible. There are countervailing trends that are preventing large segments of the population from seeing information that they may need. Some obstacles are the result of the commodification of data, others come from the technical complexity inherent in these tools. Archival materials are being swept up into a bewildering array of database formats just as newspapers, journals, and books have been.

Tomas A. Lipinski has done groundbreaking work on the growing commercial threat to the right to information as codified in the Universal Declaration of Human Rights. Access to information is a critical need in a democratic society and should be guaranteed to every citizen. The

trend toward a proprietary ownership of data and the right to information are clashing. Lipinski perceives that the former has the advantage: "Commercialization of information is gaining 'juristic and ideological ascendancy.'"³⁷

One cannot blame an industry for exploiting a marketing niche that provides tempting profits at the expense of the values of a profession. It is up to that profession to structure the negotiations with commercial vendors in a more appropriate way. But how? Commodification of archival materials presents a special problem that needs to be carefully considered. Commercial vendors that provide online access to journal articles, newspapers, and books charge heavy subscription fees that only large libraries can afford, and negotiate restrictive licensing agreements that limit the people eligible to use them. A large university library, such as the one at Stanford University, may have subscriptions to over seven hundred licensed electronic databases for locating journal articles and other research information. These are package deals with overlapping coverage: some materials are duplicated, other runs of periodicals have large gaps. For many journals, comprehensive access to the entire run requires using a patchwork of print editions plus microfilm, microfiche, and digital surrogates. Different titles require piecing together the complete run differently. The expense limits access to members permitted under the licensing agreement, thus hampering the ideal of egalitarian availability of information. The ad hoc free market history of these services means that they are incomplete, incompatible, and often awkward to use, thus hampering intellectual access.

The market forces are hard to resist. Some research libraries are spending more acquisitions dollars on databases than on print materials; sometimes the electronic version of a journal is priced higher than the print version. Several large companies have been marketing digital archives. Industry giant Google, famous for digitizing books, has looked into large-scale scanning of primary source materials. As these companies assimilate archival sources, the same issues emerge.

In commodifying print materials, database vendors do not have an absolute monopoly. In most cases, widely distributed print versions of

the articles are available at public libraries or through interlibrary loan agreements. The main advantage provided by the commercial vendor is speed and ease of use. When a vendor acquires exclusive rights to archival materials that are unique and one of a kind, the public does not have a readily accessible alternative source. When the commercial licensing model is used for archives, even the purchaser does not have the materials in any permanent form: the data goes away as a result of a lapsed payment, shifting company ownership, or changing company policy. Some vendors of online journals have been known to remove articles from the database. During economic downturns, libraries are forced to cut back on subscriptions. With print there are back issues on the shelf. With online products there is the definite risk of having nothing to show for previous payments unless the contract was carefully negotiated with that eventuality in mind.

Archivists interested in preserving equal access to primary sources in the digital environment need to consider what has happened as academic journals went from print to digital format. Database vendors charge what the market will bear. For some of them, the price is negotiable and they do not want one institution to know what another is paying. Information on pricing is typically not available on company websites. One outstanding source for journal articles is JSTOR, which started as a grant-funded database and is now a not-for-profit venture. The subscription cost for the nonprofit JSTOR database is still too high for many libraries. Other for-profit vendors charge even more. For quality products, subscriptions costing twenty thousand dollars per year are normal; some subscription costs can reach one hundred thousand dollars per year. In ten years, such a subscription becomes a million dollar acquisition—and the library does not even own it, but is essentially leasing it on a yearly basis. Often there is a cost for the back file as a one-time payment, then an additional yearly cost. Few middle-tier institutions can afford annual subscriptions in this range. Even if one hundred students accessed the database yearly, the cost per actual use could be in the thousands of dollars. Many research libraries are reaching a tipping point in that more than half of their acquisitions budget is spent on digital sources. Digitized sets of archives are among them.

Increasingly, the text of laws, legal decisions, and building codes are only available from commercial vendors for a hefty fee. This model is not inevitable. The European Union uses EUR-Lex, which provides direct and free access to all European Union law. There are some grassroots efforts to provide American laws in an "open source" mode, but so far the progress has been slow. Commercial interests have been allowed to monopolize the market. With archives, there has already been some loss of control over microfilm and digital surrogates, but since the process is at an earlier stage, there may be time to reestablish a better balance of interests.

This pricing of information has created three tiers of libraries: the haves and have nots, and those struggling in between with minimal resources. Even within a well-funded university, certain high-end legal databases are made available only to law faculty and law students. Humanities students, paying hefty tuition for the privilege of attending the school, may want to conduct research on a legal topic and be denied access. Members of the general public and unaffiliated scholars are completely excluded from the information club. Independent scholars who gain access to the rare original documents at the Bodleian may not be able to use the online databases because of the licensing restrictions. The promise that digital tools would expand access to archives is threatened by the profitable marketing models from the world of academic journals.

The business models for these information providers are constantly evolving. The companies are bought, merged, and sold on a regular basis, jeopardizing the continuity of their product. These businesses have provided valuable services, but it is up to the customers to keep them customer-oriented. That requires constant monitoring and quality control. One resourceful online company repackaged the chapters of out-of-copyright books as new articles and was able to market this bogus online resource to a large number of university libraries that were purchasing by blanket order without much oversight. Price gouging began with new marketing strategies for scholarly journals in the 1980s. Academics researched and wrote articles to win tenure, and typically were supported by their university salaries while they did so. They were delighted to see their research in print. (It is a rare academic who gets paid for a journal article.) Then

the publishers print and distribute the journals, selling them back to the very institutions that provided the free intellectual labor. Over a period of about a quarter of a century, the price structure for journals exceeded inflation by a wide margin, and journals became a bigger cost than books in acquisitions budgets. The same marketing strategy can be applied to the scanning and distribution of digital copies of archival materials. Vendors and corporate support play a key role in the research community, but it is a matter of defining the right role and managing the right balance.

And the digital sources sometimes inexplicably disappear from the Internet. One example is the case of the "Paper of Record," a digital archive of early newspapers including rare Mexican newspapers. It was being used avidly by historians, happy to avoid a long-distance trek to sort through crumbling newsprint. Even briefly using documents printed or written on brittle old newsprint paper can result in a distressing scene: a library table and floor covered with the crumbs of history. The digital version was a welcome improvement for both access and preservation. Then it just disappeared. It was quietly purchased by Google and taken offline. There were plans for it to reappear, but at any time the company can impose a high price for this now monopolized and commodified resource.³⁸ In another case, the Research Libraries Group (RLG) created the Cultural Materials Initiative, a subscription database of digital surrogates of manuscripts and other cultural objects from dozens of institutions. In 2007 the database was cancelled when RLG merged with the Online Computer Library System (OCLC). These cases demonstrate the fragility of the digital archive and the implications for access.

Increasingly, information has become commodified and assigned a price in the marketplace. Even nonprofit providers may charge hefty fees and make decisions based on financial rather than cultural values. The government-provided database PACER, which provides access to legal documents from the courts, requires a credit card to cover access fees that may be higher than the cost of providing the service. The main obstacle is no longer a gatekeeper physically restricting access, but a high price tag that ensures that only the wealthiest institutions get a license.

Large corporations have contracted for exclusive rights to market surrogate copies of archival collections. Microfilm sets of archives are already prohibitively expensive, not just for individuals but also for many institutions. The author is familiar with one microfilm project where the product was so expensive, in the half million dollar range, that fewer than ten libraries around the world could afford a complete set. In theory a competing company could rescan the documents and sell them competitively at a lower price. Realistically, no archival administrator would want to subject a collection to the stress of scanning more than once.

Google has famously experimented with digitizing both book and archival formats on such a huge scale that some observers worry about the monopolization of digitized materials. Google's management seems on the whole to be very enlightened, and interested in broad access. We know from the experience with digital journals that once a monopoly on digital assets has been established, the temptation for price gouging will be difficult to resist. There are many companies trying to get a foot in the door. It can be very tempting to allow a commercial venture to do the scanning in exchange for a digital preservation copy. When that arrangement creates an exclusive and pricey product, it may be worth rethinking the most ethical access strategies for a nation's cultural property that should belong to its citizens.

With paper archives, faculty can visit the institution that holds them and work there. It requires time and travel funds. With commercial document databases, the information may be as close as a reading room with WiFi, but access may require affiliation with the possibly elite institution that bought a restrictive licensing agreement. Instead of the secrecy and restrictions of traditional archives, the obstacles to access come from the commodification of data, including the information derived from archival formats. The open and equal access needed for a healthy democracy is impeded just as successfully by these financial obstacles as by any physical barriers.

As technology evolves, access requires knowing how to reformat databases for accessibility, and also knowing how to navigate legacy databases. Migrating data is a well-known problem. The rapidly changing software

and hardware platforms, a form of planned obsolescence, ensure that sales are robust, but the result is much lost data as systems crash and are replaced by new models. Even the technical wizards at NASA discovered in 1999 that they could not read digital files from the 1975 Viking space probe because of obsolescence. For very valuable digital objects, such as early photographs of outer space, it may be cost effective to retroactively invent drives to read the obsolete formats, a digital preservation process called "emulation." But normally the cost is prohibitive. Outmoded formats have long been an issue: just try to find a Dictaphone to play a recording belt from the 1960s. Increasingly the problem will escalate until a common ground for compatibility standards is established.

One unsung hero is retired NASA archivist Nancy Evans. Her first contribution came in 1986 when she recommended the preservation of images from the 1966–1967 Lunar Orbiter in climate controlled storage. This was the era when a great deal of the fragile digital imagery from space was being lost through neglect as images degraded over time and equipment for reading them was discarded as obsolete. Evans knew that the preserved, but obsolete, two-inch Lunar Orbiter magnetic tapes would be unreadable without the specialized tape drives, most of which were being destroyed. Evans had the foresight to rescue four of the devices, each of which weighs half a ton and is the size of a refrigerator. She simply stored them at her home. Now they are being refurbished to bring back images from the moon landings.³⁹ Certainly nothing in the code of ethics would require an archivist to store large pieces of obsolete equipment at home for two decades. The formal rules did not help; only the deeper ethical values of the profession provided the context for her decisions.

One point needs to be clearly recognized: attempts by industry to protect products from competition by creating artificial incompatibility have played a major role in obstructing free use of information in all media and across all frontiers. Computer products come in a bewildering array of formats that effectively diminishes equal or equitable access. One savvy researcher will work from his office Internet connection and get full text documents immediately with a few clicks. Another will waste a day's valuable time trying to get past passwords and incomprehensible instructions.

At present we are stuck in a kind of technological Darwinism; it is the research survival of the technologically fittest. Those researchers who catch on, get to the text; those who are more focused on content than on database architecture are left on hold listening to classical music while they wait for tech support to answer the telephone. And to take the Darwinism example a bit further, no one researcher, however brilliant, can possibly learn all the quirks of all databases. It takes start-up time even for the computer adept, so researchers learn some strategies and formats and not others, something like the evolutionary specialization that Darwin found in the Galapagos Islands.

Interdisciplinary research is heavily hit by this process of specialization and compartmentalization. Frequently even a single database will have more information than a typical researcher has the savvy to unlock. Some historians can use some functions and others rely on different aspects. The widely differing interfaces can be very confusing even to the most skillful scholars, who can be expected to master a few of them but certainly not all. This is sometimes referred to as "feature shock"—too many features to master. Researchers should not be expected to spend their valuable sabbatical time learning new interfaces as the familiar ones become unusable due to planned obsolescence.

As the situation stands, digital databases, including archival ones, are plagued by five serious flaws:

- exclusive licenses
- incompatible technical architecture
- rapidly changing software
- impermanence
- high cost

All five characteristics probably enhance profits, but they are incompatible with ideals of open and equal access. There is certainly an important role for commercial vendors, but these corporations need to serve the information community, not the other way around. Can there be an alternative

that recaptures the old nineteenth-century ideal of a noncommercial, free learning commons?

Can Archives Be “Open Source”?

The Internet is fundamentally an access tool. Reference archivists, often erroneously regarded as rather stodgy, immediately began to pay close attention to the potential of the web, as attested to by the topics that began to show up at SAA conferences starting in the 1980s. They saw the connection between instant access to government records and transparency in a democratic society. Archivists and librarians have long espoused a lofty faith in the power of knowledge to support a just and well-functioning society. The goals continue to be open and equal access to the information and evidence contained in archival materials. Archival theorists, including Margaret Hedstrom, Paul Conway, David Bearman, and many others recognized the ephemeral quality of digital documents very early.⁴⁰ It is worth looking at their list of publications and papers to see how much serious attention has been given to the issue even as sources are being transformed faster than anyone can fathom or control them.

Then and now, the major complaint from archivists has been the lack of infrastructure and standards. The technology marketplace has been either reluctant or unable to construct simple standards for compatibility and interoperability. The archival profession is dependent on commercially developed products. In the short term, incompatible components provide companies with a competitive advantage. In the long run, it hampers the development of a coherent information policy.

Daniel V. Pitti, who pioneered the Encoded Archival Description standards, is fond of quoting Charles C. Jewett's 1853 report on constructing library catalogues: “Now, even if the one adopted were that of the worst of our catalogues, if it were strictly followed in all alike, their uniformity would render catalogues, thus made, far more useful than the present chaos of irregularities.”⁴¹ The situation in the twenty-first century can once again be described as “the present chaos of irregularities.”

The 1920s were probably the golden age of standardization as the Department of Commerce worked with industry to produce voluntary uniform sizing for manufactured products from bricks to bedsprings. It was in the 1920s that the Bureau of Standards worked with the U.S. printing industry to establish 8 ½ x 11 inches as the standard letterhead size.⁴² This simple, freely adopted standard created a wave of increased access to information in ways the original Committee on the Simplification of Paper Sizes in 1921 could not have imagined in their wildest flights of futuristic fantasy: typewriters with a uniform width for the carriage, compatible fax machines at distant locations, and computer printers that print 8 ½ x 11 inch pages. In the archives, this simple standard, which originally had something to do with the size of Dutch paper molds of the seventeenth century, enabled the use of manuscript boxes in mostly uniform sizes filled with uniform folders. In the digital realm, such consistency to eliminate waste and facilitate accessibility remains elusive, despite many determined attempts.

The proliferating archival databases have evolved in an unregulated free market environment without the kind of uniformity that was still possible in 1921 when paper size was standardized. Archivists have certainly attempted to rationalize access to online information with standards. Already in the 1980s RLG and a group called the National Information Systems Task Force (NISTF) launched the uniform standards for collection-level description called AMC, based on the Machine Readable Cataloging (MARC) format. The MARC AMC format worked well for the collection-level description, but it could not be expanded to accommodate finding aids, which average twenty to thirty pages, and can run to thousands. And MARC AMC was “flat,” unable to track the nested hierarchical relationships in archival registers.

IT professionals collaborated with archivists to create EAD, Encoded Archival Description, an invaluable tool for creating compatible registers with online file-level descriptions, as developed by the Berkeley Finding Aid Project. Daniel V. Pitti, recognizing the need for an access tool independent of software and hardware, took advantage of the newly available Extensible Markup Language (XML), a simplified version of the

Standardized General Markup Language (SGML). The alpha version of EAD was released in February 1996 and rapidly became a national standard with international participation. It soon became possible to search through thousands of finding aids at hundreds of institutions. In 1998 the Online Archive of California (OAC) encompassed 2,697 EAD finding aids. At that point there were no discrete text objects, but already 168 finding aids contained embedded digital images. By 2009 some 150 separate institutions had contributed 11,840 EAD finding aids to the OAC; these finding aids included 179,209 discrete image objects and 10,846 discrete text objects.⁴³ One relatively simple break-through idea opened up a universe of access through compatible finding aids. It was a stunning success.

What began as an online set of local finding aids turned into a state-wide union catalog with links to digital surrogates. From the beginning in 1994, Daniel Pitti envisioned the finding aids as gateways to digital objects, either as full text or images. It was to be an expandable access tool, one that would both increase use and save wear on the original. The goal was enhanced preservation linked to enhanced access: "by making surrogates of the most used portions of our collections available, we can simultaneously increase access and limit physical access to endangered collections."⁴⁴ Fifteen years and hundreds of thousands of searches and "hits" later, that goal is being realized. It will be interesting to see how much full text is eventually made available for unmediated use and how much material will only be cataloged and still require a visit to the archives reading room. Researchers, of course, much prefer the unmediated direct access. The obstacles are both legal (copyright and privacy concerns) and technical (coping with the complex and expensive scanning, coding, and metadata requirements that are often beyond the means of the contributing archives). Each contributing repository decides on its level of involvement. The rate of growth will vary depending on economics, but the direction is clear. Researchers are closer to their dream of sitting in their offices and accessing unique archives from scores of geographically dispersed locations.

The integration of finding aids with full text digital documents as available on the Online Archive of California constitutes a union catalog in the fullest sense. All twelve thousand finding aids from 150 different archives

use the same interface; anyone, anywhere in the world, can access these digital primary sources. There are no user application forms, no letters of reference, no reader cards, no user fees, no subscriptions, no exclusive licenses, no passwords, no oath. And it works. But it is expensive and vulnerable in times of economic stress. It is not clear whether or for how long the system will remain free of charge.

Initial funding for development came from University of California internal grants. The grant money enabled the pilot project to provide mark-up services for repositories that did not have the technical knowledge to do it themselves. As grant money was exhausted, the technical side had already become easier with the use of templates and boilerplates. As funding permitted, smaller repositories outsourced the scanning and mark-up chores—a proper use of commercial services—and the repositories still retained control over the archival sources. Using the same principles, many other regional union catalogs have formed both in the U.S. and Europe. EAD, METS, PREMIS, DACS, Dublin Core, and the Archivists' Toolkit are all highly useful initiatives in creating interoperability. Both the Library of Congress and the National Archives have invested heavily in the search for solutions.

Some of the pioneers in providing online access to archival finding aids in a fee-free environment include the Online Archive of New Mexico, Texas Archival Sources Online, Kentuckiana Digital Library, North Carolina Encoded Archival Description, Northwest Digital Archives, Virginia Heritage, and Rocky Mountain Online Archive. The British have developed A2A (Access to Archives). MALVINE covers Manuscripts and Letters Via Integrated Networks in Europe. There is also the Archives Hub. It is an impressive accomplishment in less than two decades.

The Library of Congress has more than nine million digital manuscripts and other multimedia objects on its American Memory Website. The National Archives has created two systems, one for finding aids (Access to Archival Databases), and one for digital documents (Archival Research Catalog). Keeping track of all the websites is a formidable task, which requires a free subscription to a website evaluation site such as the Librarians Internet Index or the Public History Resource Center or sites

such as ResearchBuzz.⁴⁵ There are simply not enough cross connections between online primary sources. And there is a disconnect between the digital scanned paper item and the description and identification of that item. The good news is that integration is feasible. It does take political will and the ability to negotiate the right relationships among commercial, nonprofit, and public institutions.

We do have a workable open-source model to compete with the commodification of information. The ongoing challenge is to work individually and as a profession to integrate commercial and not-for-profit initiatives in a way that takes advantage of the strengths of each for the purpose of better utilization of data. It is a moving target. As these access tools grow, they become very expensive to maintain.

The fact remains: these well-crafted and well-funded efforts have not tamed the proliferating and sometimes deliberately confusing array of access tools, inconsistent coverage of archival sources, and confusing vendor-provided interfaces that serve commercial purposes more than the principle of access. The vigor of the professional databases is periodically threatened by the unpredictable economic crises that hit nonprofits and universities particularly hard. It makes an archivist nostalgic for the days when the weighty red books known as LCSH, issued by the Library of Congress with uniform subject heading lists, organized knowledge for the nation. Their reassuring presence in the reading room guaranteed that any researcher could find the keys to the kingdom. No more.

No reference staff can solve these technical access problems directly. Instead of confronting traditional restrictions imposed on paper documents by a protective donor or institution, we confront the practical restrictions imposed by a complexity that simply takes too much time or money to master. Must we delete article 6 of the SAA Code of Ethics?

There are two fronts in this war. One consists of small meliorations on a case-by-case basis by enlightened staff. The other is a united advocacy role by the profession as a whole to recommend better industry practice for the benefit of research. No one knows the researchers' problems better than the reference archivist and no one is better able to aggregate the experience in a way the individuals themselves cannot. Reference archivists

are reinventing their traditional role as mediators in this space between chaotic data and the information seeker. From personal experience, the author senses that the motivation for taking on such a huge task is highly idealistic, and tied to the ethics and core values of the profession.

What are archivists doing to maintain equal and open access in the online environment?

- Archivists are essential in identifying aging digital data at risk and recommending preservation reformatting. The reference archivist knows what collections are in demand. While ideally all electronic data would be scheduled for migration to current formats, it is not a bad fallback position to prioritize those databases most in demand. Researchers will find problem areas that are not obvious from the finding aids. Reformatting researcher-identified obsolescent formats is a pragmatic way to ensure that the materials most in demand are the ones that receive funding for treatment first. Reference archivists supply a valuable service by assisting in this area to ensure that needed information is not lost. These are small ameliorations that add up. But more global strategies are also required.
- On a larger stage, there is a responsibility for creating a trustworthy infrastructure of reliable standards. Ethical reference archivists find themselves working as the advocates for researchers. The foremost task in the current era is the construction of technical tools for integrating digital documents in a user-friendly format available free on the Internet—systems with an emphasis on accessibility and authenticity. The UCSF Tobacco Control Archives form a model for topic-focused full text collections, also available free of charge. The Online Archive of California serves as a model for open, equal, and free access to finding aids.
- The venerable values of public service, long practiced in archives reading rooms, will need to be expanded and redefined for the virtual reading room. If researchers are finding archives on the Internet in their offices, reference service needs to find its way into

offices as well. The Tobacco Control Archives placed millions of pages of full-text documents online. In 1995 about two thousand users accessed the equivalent of four million pages from the cigarette papers. "Internet reference use has been enormous, and it would have been impossible to successfully meet the demand in a traditional, supervised reading room environment where staffing is minimal."⁴⁶ Distributed information demands a virtual reading room with assistance provided by email, text messaging, and blogs. If reference services do not make this transition, the functioning of the information society will be distorted.

- Championing the old public library model of open access to information free of charge will continue to be a major challenge in the foreseeable future. There is an opening for steering new development away from fee-based licensing of information. Commercial ventures are sensitive to public opinion and can be influenced to provide free access to basic information, or to keep use fees within a reasonable range. One example is the opening to the public of OCLC's catalog utility World Cat on the Internet in 2006. World Cat has expanded its entries to include books, articles, archives, manuscripts, and multimedia.⁴⁷

A highly useful volume on the subject, *Archives and the Digital Library*, edited by William E. Landis and Robin L. Chandler, is peppered with the usual alphabet soup of acronyms typical of any writing on technical subjects: On one page alone the reader encounters MODS, MARC, METS, CDL, ICT, JARDA, MOAC, XSLT, URL, HTML; on the next page CSS, CMS, XTF, DLF, CMIG, DLXS, EAD, ARKS, GenDB, OAC, MOA2, etc.⁴⁸ In the midst of the tech speak in this volume are words like trust, authenticity, disclosure, reliability, stability, fidelity, integrity—words with ethical content. What is a trusted repository? What is object integrity? Old-fashioned Sunday School virtues are suddenly emerging in the midst of highly technical discussions of digitization: "for the digital repository, trust involves scholarship, authenticity, reliability, and persistence over

time and has little relationship to immediate financial rewards.”⁴⁹ These are moral and ethical concepts in the development of archival databases.

With the concept of the trusted digital repository, the ethical ideals have reached a new level and a new challenge. Historically the first phase in archival practice was veiled in secrecy and privilege. The trusted medieval archivist protected his patron's secrets from prying eyes. He kept Cromwell from removing a manuscript from the Bodleian Library. The Enlightenment-era archivist ensured the authenticity of the public record for the greater good. Then, gradually, the democratic principles of open and equal access began to take hold and slowly became established in theory and practice both in the U.S. and Europe. Archivists were at the forefront of the movement to provide direct and free access to information. Library and archives codes and access policies were adamant on the subject. Late twentieth-century archivists worked toward a model of transparency and equality. The postmodern archivist has a more complex challenge: preserving a sense of trust in the face of massive change. Business plans to monetize the data in archives offered both great opportunities for improved access and at the same time the threat of expensive and exclusive “gated communities” of information. The digital revolution provided huge profit incentives for commercializing journals, then books and archives. The information profession failed to bring the vendors' skills and resources “under the tent,” to utilize the business models in ways that would support free research. The open source movement is countering that development, but as yet there is no equilibrium. New ethical codes need to address this issue. The profession as a whole needs to formulate a twenty-first-century version of the successful public library movement that began in the nineteenth century and flourished in the twentieth. The charge is to make large quantities of data open, available, and usable. Commercial tools need to be rationally structured in the service of open and equitable access to all media, across all frontiers. Just as the equal and open access policies of the of the recent past fueled creativity and innovation, freeing digital archives of excessive financial obstructions and licensing restrictions will undoubtedly open up entire new fields of inquiry, and take learning and scholarship in new directions.

Learning from failure: The case of the disappearing Web site

Francine Barone, David Zeitlyn, and Viktor Mayer-Schönberger

Abstract

This paper presents the findings of the *Gone Dark Project*, a joint study between the Institute of Social and Cultural Anthropology and the Oxford Internet Institute at Oxford University. The project has sought to give substance to frequent reports of Web sites “disappearing” (URLs that generate “404 not found” errors) by tracking and investigating cases of excellent and important Web sites which are no longer accessible online. We first address the rationale and research methods for the project before focusing on several key case studies illustrating some important challenges in Web preservation. Followed by a brief overview of the strengths and weaknesses of current Web archiving practice, the lessons learned from these case studies will inform practical recommendations that might be considered in order to improve the preservation of online content within and beyond existing approaches to Web preservation and archiving.

Contents

[Introduction](#)

[Research methods and process](#)

[Beyond link rot](#)

[Typology of sites gone dark](#)

[Case studies](#)

[Sites at risk](#)

[Discussion](#)

[Current practices and perceptions](#)

[Conclusions](#)

[Recommendations summary](#)

Introduction

Conducted during 2014, the *Gone Dark Project* has investigated instances of Web sites that are no longer online and which have not been captured by the Internet Archive or other archiving initiatives. We wanted to examine what has happened to Web sites, valuable archives and online resources that have disappeared, been shut down, or otherwise no longer exist publicly on the Internet. Web archiving services, including national libraries such as the British Library and U.S. Library of Congress

as well as non-profit organizations like the Internet Archive, are dedicated to storing the contents of the Web and have had great success in preserving online content as part of recent human history (see, for example, BBC News, 2010; Lohr, 2010; Internet Archive, 2014). Despite these efforts, however, some important content has not been archived. Other research (cited below) shows that the lifespan of online content is pitifully short. The average lifespan of a Web page is difficult to determine, but estimates put it at a mere 100 days in 2003, up from just 44 days in 1997 (Taylor, 2011; Barksdale and Berman, 2007). As the Web evolves, different types of content become more susceptible to loss. In 2008, a survey by blog search engine Technorati found that a whopping 95 percent of its 133 million tracked blogs had been “abandoned” or not updated in 120 days (Quenqua, 2009). Ironically, in May 2014, Technorati unceremoniously shut down its famously extensive blog directory — once an indispensable tool to the blogosphere — with no prior announcement and little to no media coverage (Bhuiyan, 2014).

A study published in 2012 (SalahEldeen and Nelson, 2012) reveals that historically significant social media content decays at an alarming rate with 11 percent of timely media content lost within one year, rising to nearly 30 percent in two years (at a rate of .02 percent of shared resources lost per day). Compounding the problem of disappearing Web sites is the issue of link rot or hyperlink decay. In a study of academic references, Zittrain, *et al.* (2013) found that over 70 percent of URLs in academic journals and 50 percent found in U.S. Supreme Court opinions have broken or no longer link to the original citation information.

These losses and “gray areas” on the Web suggest that between what is automatically crawled and saved and what becomes lost without much impact in day-to-day activities on social media, lies a large swath of the Internet that we know very little about in terms of historical record. Continual reports of missing Web sites and 404 errors suggest that there are still Web pages that “go dark” on a regular basis. In fact, despite Google’s Cache, the Internet Archive’s Wayback Machine [\[1\]](#) and national digital preservation initiatives, it is still easy for a site to be completely lost from the public Web. Is there significant data loss in these situations? Even Web crawlers that have captured billions of pages cannot save all the content from sites before they vanish, especially if those sites are not widely known and/or indexed by major search engines or if the content is held in a database inaccessible to crawlers. There may be cases where the data is still held privately or off-line where Web crawlers cannot find it. Can it still be recovered?

In the light of this, the Gone Dark Project wanted to address the concern that there may be instances of culturally valuable Web sites which are no longer online and whose disappearance represents a major public or social loss. What, if anything, can be done to mitigate future losses of this kind?

As a collaborative project between Oxford Anthropology and the Oxford Internet Institute, this project benefited from both a technical and anthropological approach to the subject of digital content loss. We were able to investigate actual cases of content loss on the Web, including interviewing the original content owners or other involved parties, in order to better understand current practices and inform future innovations in pragmatic Web preservation.



Research methods and process

The Gone Dark Project was conducted over nine months from February to October 2014.

The questions guiding the research were:

1. How significant and/or widespread a problem is the disappearance of Web sites?
2. What common factors result in important Web content not being archived?
3. What practical steps or changes to Web preservation practices and/or policy can be identified to mitigate against reoccurrence in the future?

We should clarify that our principal concern was with sites which contain substantial or significant content, rather than either social media posts or collections of links to other sites. We explain this in more detail below.

The first task undertaken was to identify as many cases as possible of such sites known once to have existed, but which are no longer publicly available online (especially those that are not well-archived in some form or other). This allowed us to gauge the scope of the problem of Web sites 'going dark'. It also brought up methodological challenges; notably, how to find digital artifacts that no longer exist by looking for clues around the Web.

Searching for ghosts of Web sites required pragmatic methods that evolved over the course of the research. For example, creative use of search engine filters and existing archive resources such as Google's Cache and the Internet Archive's Wayback Machine was essential. While reference was made in the initial scan to previously compiled lists of dead or endangered Web sites, apps and services (*e.g.*, ArchiveTeam's DeathWatch [2]) these were largely of limited utility because most sites on popular lists were deemed to lie outside the scope of the Gone Dark Project (see below).

Potentially relevant sites or directories of pages were scanned for dead links using automated link checkers to find broken references to databases, repositories, or archives of original content. We also conducted some manual trawling through lists of links on older sites, including academic indexes. These links were then followed up to collect background information about the nature of the site, reasons for its disappearance and whether a public archived copy exists. For paradigm cases attempts were made to contact the relevant site owners and interviews were conducted. We wanted to learn about what happened to the site, how its loss might have been — or might still be — avoidable; and to trace the current whereabouts of the original content.

We also distributed links to the project via social media platforms, academic mailing lists and user forums. We especially encouraged academics and subject experts to send us information about content-rich sites and databases that might no longer exist.

On a day-to-day basis, social media accounts on Twitter and Facebook were used to foster dialogue with individual Web archivists and large organizations directly engaged in Web preservation, including the U.K. Web Archive [3], Austrian Web Archive [4], Internet Archive [5], NDIIP (U.S. Library of Congress) [6] and Internet Memory Foundation [7], among others. Throughout the project, active engagement with specialists in the fields of Web archiving, Web history and digital humanities helped to identify case studies of sites that have gone dark as well as to better understand the processes and professional standards for crawling and archiving the Web that are currently in place around the world. Informal surveys and interviews with social media followers were fruitful in pointing out strengths and weaknesses in current practices. The social media channels greatly informed the data analysis and final recommendations of the project.



Beyond link rot

Many of the reports of sites no longer available are attributable to 'linkrot'. In these cases, the original or referring URL no longer works (for many reasons) generating a 404 not found error. However, many sites whose published URLs no longer work do still exist, but at other URLs (site restructuring or redesign may break many links, as clearly needs to be pointed out to Web designers and the site managers who employ them). Even where the original site is no longer available, its content may have been preserved through one of the many initiatives to archive Web content. There are cases, however, as we shall see, in which the content in question has not been captured. These often involve sites where content is delivered by a user-searchable database such as a catalogue.

One such case in point is the Haddon catalogue developed by Marcus Banks in Oxford with support from the U.K.'s Economic and Social Research Council (Grant R000235891 [8]). The project sought to document early (pre-World War II) ethnographic films. In the course of the research, some 1,000 instances were identified and information about them was made available via a searchable catalogue which went live in 1996. As database and server technology advanced it became unavailable: the database engine was no longer compatible with current operating systems and it went down in 2005 [9]. Since the original project funding had finished, there were no longer resources available to reprocess the data (which had been securely archived) to make it available again using a different database engine. Now some support has been offered by colleagues in Manchester, so it is hoped that access to the Haddon catalogue will resume in 2015 or 2016, after 10 years of 'darkness'.



Typology of sites gone dark

After canvassing as many known defunct Web sites as possible across all fields of interest, the second task was to categorize our initial findings into manageable types. Potentially relevant case studies were organized by theme and the primary reason for the page's disappearance. This process helped to make more sense of the wider landscape of what can be described as the vanishing Web — sites, pages and genres of content that have gone as well as are in the process of going dark, and especially those which appear to be at greater risk of doing so, either for specific reasons or simply because they have a higher rate of unintentional decay.

Main types of sites:

1. Scientific and Academic: Databases, research tools and repositories ranging from the natural to social sciences and humanities. Losses of this type are commonly the result of the end of funding or institutional neglect, in which case the original data may still be held (*e.g.*, on university servers).
2. Political: Personal homepages of politicians, campaign pages, political speeches and/or repositories of once-public government files. Some journalistic sites also fall under this category (we discuss a case below).
3. Historical and Cultural: A range of sites with different origins fall within this category, including collated collections of historical documents, genealogies or research portals, as well as more professionally run film, video or music archives.

4. “Labours of love”: Specialized project pages or information aggregation sites, typically self-hosted and curated by independent individuals with little to no institutional backing.
5. Social media: These include popular Web services on sites run by companies such as Google [10], Yahoo! [11] or Microsoft [12], including blogging platforms, social networking sites and other utilities that change hands or have been retired since social media platforms and startups evolve quickly and come and go easily, often leaving behind users with data they would prefer to keep (examples include several popular Web services from the late 1990s).

Main reasons for sites disappearing:

1. Neglect: Intentional or unintentional neglect is probably the most common reason that a site disappears, including allowing domain registrations to expire; losing or not updating files; and not keeping adequate backups.
2. Technical: Technical issues are usually bundled with some form of neglect or insufficient financial resources. Purely technological reasons for content loss include hardware malfunction, viruses, Web host errors and accidental file deletion.
3. Financial: A common factor among sites gone dark is the cost of site maintenance (hosting fees and/or server maintenance, plus staff costs where relevant.)
4. Natural disaster: Computer hardware is susceptible to fires, floods, rioting and neglect (just as are paper files). Although in principle “Lots of Copies Keeps Stuff Safe”, for many reasons the many copies may not have been made or distributed.
5. High-risk situations: Tumultuous political climates are a nightmare for data loss. Sites can be shut down intentionally by hostile regimes or otherwise lost during human rights crises. Legal prosecution or the threat of this can lead to the removal of material: the international legal saga about the availability of material to do with the Church of Scientology is a case in point [13].
6. “Web wars”: Competition between top Web companies such as Google, Yahoo!, MSN and AOL leads to aggressive acquisition of popular services that are subsequently abandoned, shut down or absorbed into a larger platform.

An anonymous *First Monday* reviewer points out that the interconnections between neglect, financial constraints, and technical issues are particularly insidious. A sort of fatal creeping obsolescence can occur that is caused by a mix of under-funding, lack of investment in technical updating and neglect that is very different from a simple site crash or attack that exposes that the backups had not worked.

On the whole, it was clear that a) some sites are going dark across the Web without being archived and b) those sites vary widely in size, type and content. This confirmation is in itself a significant finding. However, the main aim for the Gone Dark Project was to focus on sites of particular socio-cultural value that constitute an irretrievable loss notably marked by large amounts of content not likely to be saved/crawled by automated software. Abandoned blogs, deleted user profiles and short-lived Web app startups are all digital losses, yet they have largely become accepted and even expected within the landscape of the Internet today. As the Web evolves, people move on and leave a patchy trail of online interactions in their wake. The dynamics and ethics of digital preservation of content like personal social media postings remain debatable and outside the scope of this paper (see Mayer-Schönberger, 2009).

While each and every site that goes dark arguably constitutes a lost piece of Internet history, the case studies chosen for deeper investigation were selected on the basis of being of cultural, heritage or

social value whose loss represents a cautionary tale for Web preservation. There is certainly a considerable element of personal judgment about what constitutes an ‘important’ Web site containing ‘valuable’ material. Recognizing this, we would also observe that this is by no means a new problem: all archivists have always had to make decisions about what to include and what to reject from inclusion in the archive under their control. These judgments (albeit individually questionable) often include assessments of what future researchers, ‘users’, or lawyers might find helpful. The decision to archive may not be clear cut in any one case, but the intuition behind it remains clear. As one research discussing a film archive has it, an archive is a bet against the future — betting that these records will be found useful [14].

Once categorized, a more narrow focus was taken for the remainder of the project. The following section will focus on a selected number of illustrative cases of sites gone dark, including what happened to the data, if it still exists; and to interpret how each case can inform recommendations for future prevention. Rather than simply collect lists of defunct and unarchived pages, the Gone Dark Project sought out the original content owners in order to discover the individual stories behind the 404 error page, or, more simply, to find out what happens when a Web site dies.



Case studies

The selected case studies below illustrate various ways in which valuable Web sites can go dark. The first two examples represent instances where important digital resources that were once available online have gone dark for an extended period of time. In these cases, the original content still exists, but the challenge is making it available again. The final case takes a different tack, focusing on potentially more widespread but difficult to quantify circumstances with potential impact throughout extensive portions of the Web. Together, the cases illuminate where existing Web archiving practices are insufficient due to the fleeting and impermanent nature of some Web content as well as obstacles impeding content owners’ ability to archive important data in a usable format before it is too late.

Kwetu.net

Kwetu.net [15] (“our home” in Kiswahili) was a privately owned grey literature site established in 2000 by Karani Nyamu and Luke Ouko as an online repository of photos and videos from Kenya, Uganda and Ethiopia [16], eventually expanding to include content from over 30 African countries. The same year, the *Economist* had declared Africa “The Hopeless Continent” (*Economist*, 2000). The rationale behind Kwetu.net, according to its founders, was to disprove that singular narrative and counteract the dominant, lopsided portrayal of Africa as a continent of war, poverty, disease and corruption. With Kwetu.net, they intended to make accessible to the rest of the world informative documents that would showcase Africa in a more positive light.



African Resource Service

[ABOUT US](#) | [PARTNERS](#) | [SUBSCRIBERS](#) | [RATES](#) | [YOUR FEEDBACK](#) | [CONTACTS](#)



Demo Search

Login

username password

[how to use this site](#) | [faqs](#) | [sitemap](#)

What we do

KWETU (Swahili for our home), is a resource service of African content; documents, documentaries, photographs. The concentration is content on development issues dating back to the 19th Century. It is full view, allowing users to actually view and read the material online.

Contact

Flex Place, Ragati Road, Upper Hill.
P.O.Box 11444-00100 Nairobi, Kenya
Tel :254 20 272 2642
info@kwetu.net
www.kwetu.net

[Terms & Conditions](#) | [Privacy Notice](#) | [User Agreement](#)

Figure 1: A cached copy of Kwetu.Net’s login page from 15 December 2005, accessed via Archive.org.

Its mission was therefore to offer the world a wide range of African content resources — in the form of “grey” literature focusing on health, social welfare and development issues in Africa — and to provide access to it irrespective of place and time [17] (see [Figure 1](#)). The type of documents sought by Kwetu.net included unpublished reports, baseline surveys, speeches, photos, videos, university theses and a mix of historical (dating as far back as the nineteenth century) and present-day information drawn from many fields or industries — agriculture, healthcare, conservation and politics to women’s issues and urban development — that was otherwise difficult to find or largely inaccessible online [18]. As one founder put it, “if it was grey material, we sought it out”. Much of this material was acquired from PDF documents produced by education and research institutions, government agencies or NGOs. In Kenya, the founders also partnered with national archives such as Kenya Railways and the

African Medical Research Foundation to source content. At one point, Kwetu.Net had a list of 47 partners listed on their site [19].

Access was available on the following basis. The site offered a free demo/preview to all visitors, but access to the full site content was available by paid subscription only. Subsidized rates were available to African educational institutions and subscription fees for other institutions varied from US\$800 to US\$6,000 depending on the institution's size. Individuals could also subscribe for US\$50 per annum. To secure content from producers and owners, there was a services-for-content system in place. For instance, a 30 percent subscription discount was applied to institutions that provided them with content [20]. Similarly, Kwetu.net had developed a network of correspondents in over 30 countries who helped to secure new content for the time they were in active operation. At their peak, the site had a subscription base of 15 African and U.S.-based universities, according to the founders. They also served a range of other institutions such as think tanks, embassies and foreign missions, civil society and donor agencies.

In terms of functionality, the founding team built the site from scratch, including a search engine [21] and an (A-Z) index-tagging system, with a diverse range of tags. This became extremely technically demanding. Extensive amounts of time would go into negotiating with the various content producers, uploading, curating, tagging and indexing the content to ensure ease of access and searchability. Over time, the demand for more content compounded this challenge. Even when the site reached upwards of one million manuscripts in its database, it became clear that it could not supply the demands for content put on it by its paying customers. Furthermore at this point, demand was not coming from local and regional universities — primarily because of low penetration rates and high Internet costs — which stalled the spread of a localized user base.

On top of the technical challenges, the primary motivation that led to Kwetu.net going dark was financial. The founders initially established the site as a “labour of love”. Soon, the maintenance costs to keep it afloat — including paying the 30 correspondents connected to the site — exceeded the subscription revenues. The founders were therefore compelled to divert their attention to income-generating projects, and eventually away from Kwetu.net. A one-day delay in paying for renewal constituted in the loss of kwetu.net domain (www.kwetu.net is now hosted in Istanbul as a Turkish tourism site) and the team was unable to recover it.

The site officially went off-line in 2004, according to Nyamu, one of its founders. The original site including HTML, text and images is cached in the Wayback Machine, making it possible to view the skeleton of the old site. However, the search function does not work and no access to anything behind the search paywall is available, which is what made this a case of concern for the Gone Dark Project, since the paywall/database combination made the material inaccessible to the Web crawlers. Upon further investigation and interviews with the site's founders, it was possible to find out more about the large collection of data that was once available through Kwetu.net's search portal. Although it no longer exists on the Internet, the content that had been meticulously curated is still in the hands of the founders. The team, though now working on other ventures, is still passionate about what they had started and stated (in interviews in the course of this research) their interest in reviving the project.

Despite the technical and financial challenges, they do not consider the site to be a failure. So what would it take to get Kwetu.net back online? The requirements to bring the site back would be primarily financial; that is, securing enough funding to keep the project sustainable. Whether a subscription model would still function in light of the Open Access movement in academia is not clear to these writers. While founders, Nyamu and Ouko, indicate that they still have access to the content

and maintain relevant connections with their provider networks, at the moment they are focusing their attention on private sector clients.

In this case, a combination of technical, financial and human factors was involved. Referring to the list of common reasons for sites going dark presented above, at least three (neglect, technical and financial) apply here. One of the often overlooked aspects of Web preservation is the human time and energy it takes to keep Web sites alive, updated and functioning. All of the technical support fell to the original founders and immediate staff. Human oversight resulted in the domain name being lost, at which time, from the point of view of site visitors, it would have simply vanished. Finally, in cases such as this where the content is still in a state of preservation by its owners, but remains dormant due to a lack of resources, what can be done to restore it? We return to this question in our conclusion.

Europa Film Treasures

First launched in 2008 by Serge Bromberg, Europa Film Treasures (europafilmtreasures.eu) was “an online film museum”, described by its founders as “an interactive tool for the promotion of film culture” [22]. Presented in English, French, Spanish, Italian and German, the Europa Film Treasures (EFT) homepage offered “free access to a scheduling of heritage films from the most prestigious European archives and film libraries”. Online streaming of all full-length films was available without charge or geographic restrictions, making EFT an important repository and indeed a genuine public service:

Faced with the vast choice offered on the Internet and the thousands of videos of uncertain quality and often-vague origins, we propose an entirely legal film offering on the Internet. Our principal commitment is to maintain this quality cultural offering, for it seems indispensable to us that all can access it without a tariff, geographic or linguistic barrier. [23]



Figure 2: Screenshot of the Europa Film Treasures Web site before it vanished (via Wayback Machine).

The Web site was made possible through partnership between Lobster Films, Sarl [24] — a film production and restoration company based in Paris — and 31 “prestigious” European archives including those of the British, Dutch, Danish, French, German, Irish, Italian, Finnish, Spanish and Swedish film institutes, among others. Internet service provider Enki Technologies handled the software, programs and file storage. It was also supported financially by the European Union’s MEDIA program and other public and private partners. Copyright permissions to the original films for redistribution were secured from these partners, which effectively made EFT a heritage film aggregator that took on the hosting and maintenance responsibility for the films. According to the original “About” page, the collection contained 201 films dating from 1896 to 1999.

Another intention behind the site was education and instruction in the field of film preservation. All of the films, many of which were old, rare or from “relatively unknown film industries” were accompanied by an explanatory booklet of notes for better comprehension and a history of the film’s “discovery and/or restoration”. The original, full-length films together with additional film restoration resources, quizzes, puzzles, interviews and music composition notes comprised a valuable example of an interactive, public digital archive. When no music was available, Lobster would commission orchestral scores from music students (not more than one film per musician), and pay for them. Thus, the site as a whole became an important and much-loved resource with many valuable features.



Figure 3: Announcement from the EFT Facebook page regarding the site's temporary closure in June 2013.

A message from 2013 on the official EFT Facebook informed users that the site has been temporarily closed for technical and financial reasons (see [Figure 3](#)) but no specifics were given. The post promised that a new partnership may result in the site re-opening very shortly. By September 2014, no new announcements had been made on the Facebook community or any other site regarding a re-launch despite the indication that users would be kept informed. In response to this post and in other places around the Web, many former users questioned what had happened to the site and expressed their dismay that they had lost access to the videos. The full details of the financial and technical reasons for a prolonged outage have remained mostly unexplained, leaving these former site visitors in the dark.

While the text and images making up the shell of the EFT Web site have been saved by the Internet Archive, the Wayback Machine has not saved copies of the actual films. When we contacted Lobster Films' CEO, Serge Bromberg, he provided the reason for the site's temporary disappearance:

Enki Technologies went bankrupt, and before the last films went on line, we were told that the owner ... erased everything from his hard drives, and left without a trace. That was the end of the Web site as we knew it, but we of course still had the original masters for the films (on digi beta or Hdcam), the rights attached to them, and all the special contents created for the Web site. [[25](#)]

Similar to Kwetu.net and Haddon Online, the original files were luckily still in safe hands and awaiting an opportunity for restoration.

Thankfully, that time has arrived. Bromberg also informed us [26] that an as-yet unannounced new venture with Franco-German TV network ARTE will see the restoration of EFT films made available on a weekly basis via the ARTE Web site (see Figure 4). The timing of the launch of ARTE's cinema platform was serendipitous as it was an especially good fit for the EFT collection. In order to restore access, Lobster Films, backed by ARTE, covered the cost of reformatting the films for their new Web location. The Europa Film Treasures page at the ARTE Web site is currently live despite no official announcement being made at the time of writing.



Figure 4: Europa Film Treasures' new home on the ARTE network as of November 2014 [27].

The return of EFT's film collection is still a work in progress. The films will be released weekly, so the full collection is not yet available. According to Bromberg, "When all the films are re-injected, we have already decided with ARTE to keep adding more films from the European Archive's vaults, with new explanatory texts attached." [28] However, the new site is only available in French and German, the languages of the Strasbourg-based ARTE network and, as yet, the guides, educational texts and interactive materials from the old site have not reappeared.

In its current form, EFT represents a successful case of bringing a once "dark" site back online, yet lessons can still be learned regarding the dangers of digital data loss. It took the sudden actions of just one person to wipe hard drives that would take down an entire Web site for nearly two years.

Following that event, a great deal of dedicated effort, cost, negotiation and even luck went into the restoration process to bring EFT back online.

In addition, at the time of writing, the original URL (europafilmtreasure.eu) no longer resolves, so visitors to that URL or to the as-yet not updated Facebook page are none the wiser that the films are being re-released in a new location. Similarly, a search for “Europa Film Treasures” on Google (October 2014) does not yet bring up the new site. With some of the films online, but a continued lack of communication with the public both prior to and after the site was closed as well as in the lead up to its re-launch, the EFT case brings up some interesting gray zones that affect Web preservation. Even as former users made continued reports about the site going dark and that they desperately wanted restored access to the films, lack of transparency led many to assume that EFT was not coming back.

An anthropological case study approach proved effective in addressing both of the aforementioned cases. In each, there were complex organizational, financial and/or technical reasons that the content is no longer available to the public. Tracking down the relevant parties required prolonged investigation and multiple attempts at personal communication. While both Kwetu.net and EFT are the types of cases that Gone Dark researchers had expected to encounter in the course of the project, their circumstances would be difficult to pre-empt from a preservation standpoint. Each instance has localized peculiarities and complications. The benefit of both cases, however, is that the content is still in existence. The stories behind the loss of access and possible restoration can be used to evaluate what methods might be employed to restore sites like these in the future, or, preferably, to prevent such losses before they happen.



Sites at risk

Restoring a single site is a challenging enough task, but when a collection of related Web sites goes dark, prevention strategies are much more difficult to specify (and quantify) as losses can potentially include an entire digital ecosystem of information. A challenge for archivists is being able to tell the difference between isolated cases and more dispersed problems that may entirely wipe out a whole significant portion of the internet with serious social implications.

Hardware failure and technical neglect are not the only ways that online content can be lost. Most troubling in 2015 are Web resources that may be threatened by malicious parties (from hackers and militant groups to governments) who want to intentionally remove ‘conflicting’ information. The following case study focuses on conditions of political turmoil where external, and often non-digital, factors at play put the Web at risk every day. It shows that preventative backups are especially important when it is not always clear what information will become significant in the future.

Wherever there are volatile conditions on the ground, the Internet is susceptible to damage and loss. Human rights-related Web sites are therefore especially at-risk of going dark. In such cases, there are serious socio-political implications. When local news Web sites or cultural heritage organizations have their sites shut down during times of social upheaval, riots, or war, both daily communicative capabilities and the historical record can be irrevocably damaged. Unlike the two case studies above, the following study of at-risk sites in Sri Lanka from the perspective of a citizen archivist shows what can be learned from an expert who independently archives at-risk sites *before* they are lost forever.

Websites at risk

An archive of web initiatives on the peace and human rights in Sri Lanka

[Home](#) [Contact Me](#) [Digital Archives Presentation](#) [Disclaimer](#) [Sinhala and Tamil Fonts](#) [Using the archives](#)

Using the archives

There are a number of ways you can use these website archives. They are aimed at researchers, scholars and anyone interested in key websites and web initiatives with information, research and other content on human rights, peace and governance in Sri Lanka.

To access the archives:

- All the website archives are date stamped, so you can go back in time if the most recent archive does not meet your needs or you cannot find the content you are looking for.
- Download the archive to your computer and unzip them.
- Navigate to the folder you unzipped and double click on the `index.htm` or `index.html` file

COMPLETE WEBSITES ARCHIVED

[Army](#) (3)

[Berghof Foundation](#) (2)

[CaFFE – Election Monitoring](#) (1)

[Citizens Commission](#) (1)

[Citizens.lk](#) (1)

[COI](#) (1)

[Colombo Art Biennale 2014](#) (1)

[Colomboscope](#) (1)

[Cost of War](#) (1)

[CPA](#) (1)

Figure 5: Screenshot of Sanjana Hattotuwa’s Sri Lankan archive Web site, Sites At Risk.

Sanjana Hattotuwa, human rights activist and creator/curator of *Groundviews* [29], Sri Lanka’s first citizen journalism Web site, is at the forefront of endangered Web site preservation in Sri Lanka. Hattotuwa’s personal blog, Sites at Risk Sri Lanka (see [Figure 5](#)), was created as an “archive of Web initiatives on peace and human rights in Sri Lanka” [30]. Inspired by the Internet Archive’s Wayback Machine, which Hattotuwa found does not adequately archive Sri Lankan civil society content “with any useful degree of comprehensiveness or frequency”, the purpose of this site is to keep downloadable .zip copies of entire “civil society and NGO Web sites and Web based initiatives on human rights, democratic governance and peacebuilding” for when they “suddenly go off-line or are rendered inaccessible in Sri Lanka” in order to preserve the content for scholars of peace and conflict [31].

Hattotuwa’s curated archive reveals the ease at which, one site at a time, an entire ecosystem of Web content can remain at continued risk due to conflict. He reflects on why it is important not to let this happen:

The loss of digital resources for human rights activists is a significant one. The danger is two fold — one is of an enforced erasure and deletion of vital records, the other is deletion and erasure out negligence and technical failure. In both cases the failure to adequately and strategically adopt safeguards to backup information can exacerbate information loss. The issue with [human rights] documentation is that it is often irreplaceable — once lost, digitally, the same records cannot be regenerated from the field. Sometimes it is possible to go back to physical records, but most often the digital record is all that's there. [...] Digital information loss in this context can, as I have argued in the past, lead to the exacerbation of conflict. [32]

His expert knowledge of the political situation and key players in Sri Lankan human rights arena enable Hattotuwa to make pre-emptive and decisive steps towards archiving potentially vulnerable content with a higher success rate than relying on automated crawls. At the first hint of vulnerability, he saves a copy of the site in question before it can be lost. The content he is saving has a personal relevance and connection for him and he is well aware of the value of the archives of the information that he keeps.

Hattotuwa's memory of some of the greatest Web site losses he has witnessed reflects this:

The most significant loss around Web site based data in Sri Lanka I have encountered [were] on two occasions. One [was] the Mayoral Campaign of a candidate I in fact stood publicly and vehemently opposed to. His campaign team created a Web site around their vision for the development of Colombo, engendering comments for each point in their manifesto in Sinhala, Tamil and English — the languages spoken in Sri Lanka. That Web site, soon after he lost, was taken down and yet was a treasure trove of ideas around governance and urban rejuvenation. The other site loss, arguably even more tragically, was the erstwhile site of the Lessons Learnt and Reconciliation Commission (LLRC), a process that looked into stories from citizens around the end of the war. There were citizen testimonies, records of the public hearings and associated documentation on the [government's] site that was for whatever reason just allowed to expire. ... My own archive of the LLRC plus another I helped set up are now the country's only archives of this content. [33]

In terms of technical solutions, the full archive for each saved Web site is stored as a .zip file. This allows them to be opened regardless of the user's operating system. The files themselves are hosted on a public Wordpress blog using Box.net storage. Anyone can download the copies and store them locally. The .zip files are self-contained copies of the entire Web site [34], with all pages, text and images, so that once downloaded, the site can be browsed off-line as if it were a live copy, even without the need for an Internet connection. For scholars, this type of repository system has advantages over simple screen grabs or surface crawls stored on a Web server. The format means that the files in the archive can be full-text searched quickly and efficiently using desktop search software.

The site itself is functional and can instruct and inspire others to create similar collections of important Web sites that are at risk of disappearing. For instance, he chose the site name to be "scaleable": "the idea was that each country or region would use sitesatrisk and at the end plug in their name — *e.g.*, sitesatriskuk, sitesatriskkosovo" (Hattotuwa, 2008). As yet, he is unaware of any other sites replicating his model, but understands why this is unsurprising: "I am constantly responding to

some emergency or the other, constantly myself the subject of hate, hurt and harm. It's not easy, and so I understand why others haven't taken this up." Naturally, Sites at Risk Sri Lanka and all the files that appear there rely on Hattotuwa's personal dedication and upkeep. In such constantly perilous conditions, this is a difficult task to assume.

In the case of Sri Lankan human rights Web sites, Hattotuwa reveals that "a litany of issues" that are responsible for site loss, from "an incumbent regime viciously intolerant of critical perspectives on war and peace to a disturbing lack of awareness of, emphasis on and interest in safeguarding information and knowledge" by NGOs and civil society actors in Sri Lanka. What is worrying is that "most never learn, even when disaster strikes once" [35]. Thus, one solution to sites going dark is to improve general awareness of the fragility of online content.



Discussion

Who will save the Web?

Given the difficulty of tracing sites that have gone dark once they are off-line, we find that greater engagement with subject experts will be at the forefront of better Web preservation tools and practices. Since deep Web content like media or file repositories and research databases are absent from standard Web crawls (see below), *selective archiving* is best undertaken by those with firsthand knowledge of essential sites — such as career specialists, journalists, historians, hobbyists, activists, academics and private individuals. As experts typically engage in maintaining their own records of files and research repositories, they will be among the first to notice when a site goes dark and also, like Sanjana Hattotuwa, able to prevent imminent losses.

As we note below, there is a problem that it is not clearly any one person or organization's responsibility to 'archive the Internet' (the Internet Archive's self-appointed, and limited role as discussed above, notwithstanding). Outside of dedicated university or national library archive programs [36], academics in particular may find themselves becoming inadvertent archivists, unaware that the copies of content they produce in their day-to-day work may be the only remaining copies of important Web archival materials. Certainly, many digital researchers do not begin their projects intending to keep permanent archives to make publicly available. Similarly, foresight and intuition for Web preservation is not always coupled with institutional or financial stability.

A good example of selective archiving by subject experts is the Internet Archive's Archive-It program:

Archive-It is a subscription Web archiving service from the Internet Archive that helps organizations to harvest, build, and preserve collections of digital content. Through our user-friendly Web application Archive-It partners can collect, catalog, and manage their collections of archived content with 24/7 access and full text search available for their use as well as their patrons. [37]

Current subscribers include college libraries, state archives, historical societies, NGOs, museums, public libraries, cities and counties. The success of such archives to maintain important timely content

became evident when a curator from the Hoover Institution Library and Archives [38] had the foresight to include blog posts in Archive-It's Ukraine Conflict Collection [39] in July 2014 (Hoover Institution, 2014). One such blog post became a piece of contentious evidence potentially tying separatist rebels in the Donetsk People's Republic in Eastern Ukraine to the Malaysian Airlines Flight 17 crash (Dewey, 2014) that killed 298 passengers and crew. When the live post was deleted, the evidence remained for international scrutiny in the Ukraine Conflict Collection, preserved because a proactive archivist recognized its importance before it was too late.

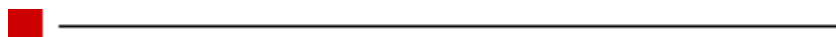
As Nicholas Taylor, Web Archiving Service Manager for Stanford University Libraries, explains:

Internet Archive crawls the Web every few months, tends to seed those crawls from online directories or compiled lists of top Web sites that favor popular content, archives more broadly across Web sites than it does deeply on any given Web site, and embargoes archived content from public access for at least six months. These parameters make the Internet Archive Wayback Machine an incredible resource for the broadest possible swath of Web history in one place, but they don't dispose it toward ensuring the archiving and immediate re-presentation of a blog post with a three-hour lifespan on a blog that was largely unknown until recently. [...] Though the key blog post was ultimately recorded through the Save Page Now feature, what's clear is that subject area experts play a vital role in focusing Web archiving efforts and, in this case, facilitated the preservation of a vital document that would not otherwise have been archived. (Taylor, 2014)

Another example from Archive-It is the Occupy Movement Collection [40], which was started in December 2011 to capture ephemeral Web content to record the then rapidly developing global Occupy movement. In April 2014, researchers decided to look back at the 933 seed URLs amassed since 2011 to see how many of the pages were still live (Archive-It, 2014). They found that while 90 percent of archived news articles and 85 percent of social media content was still live on the Web, this number dropped to 41 percent for the 582 Web sites in the collection. Fifty-nine percent of all Web sites were no longer live and either returned 404 error messages or had been taken over by cybersquatters (Archive-It, 2014). This useful analysis shows that even with selective archiving, the work of Web preservation is ongoing. Sites are still going dark.

It would be interesting to see similar statistics for live or defunct Web sites for other Archive-It collections. Automating this process might be difficult, however, as in this case, "using a human to check the URL, rather an automated process, allowed for closer analysis of the live content to determine if it was on topic" (Archive-It, 2014). Interactions with archivists at various national libraries and organizations have shown throughout this project that there is a great deal of human intervention at work throughout the entire archiving and preservation process although it might look totally automated to outside observers. Most archives accept submissions and are grateful for notifications from the public about sites that need their attention.

In the following section, we address existing archiving practices and public perceptions before offering recommendations from this research that we hope will better facilitate more comprehensive solutions to sites going dark.



Current practices and perceptions

One popular view is that nothing on the Internet is ever truly “deleted”; that is, anything we put on the Web will be around forever to haunt us (Rosen, 2010; Whittaker, 2010) because in the digital age the default has shifted from “forgetting” to preservation (Mayer-Schönberger, 2009). This outlook makes it difficult to communicate to the general public the risk of content on the Web being lost to the world. Many Internet users have become familiar with the Wayback Machine, whose mission is to copy virtually every page on the public Web. This fantastic project has both alerted people to early Web sites with nostalgic value that might have been lost as well as become an essential tool for Web site creators or bloggers to find backups of their own pages that they might have accidentally deleted. As seen above, the Archive-It service and “Save Page Now” feature also enable others to submit links to supplement the Archive. As a result, it gives the appearance that every page on the Internet is being safeguarded and therefore needs no further intervention. Indeed, in the course of interviews for this project, we encountered cases where the former owner-operators of defunct URLs simply direct previous users of their site to the archived version of it in the Wayback Machine rather than attempt to save or restore the content otherwise.

Yet some fundamental misconceptions about the Wayback Machine are employed in this reasoning. For instance, evidence from this study suggest that — in the case of sites housing significant content of cultural or social value — it is not enough to simply leave a site to be captured by the Internet Archive with no other provision for saving its content for long-term preservation. While the Internet Archive’s task to save a snapshot of the Web is useful, it is insufficient for sites that go beyond simple text and HTML.

Web crawlers like the Wayback Machine take a snapshot of surface content only. However, pages like EFT or Kwetu.net “may serve as the front end to a database, image repository, or a library management system, and Web crawlers capture none of the material contained in these so-called ‘deep’ Web resources” (Kenney, *et al.*, 2002). Contrary to popular belief, the searchable Web only represents a fraction of the pages on the Internet (Wright, 2009), omitting many types of sites and file repositories. Even where a standard crawl produces a sufficient facsimile of a largely text-based site to act as an archival record, it does not mean that it will be around forever.

Because the Internet Archive respects changes to the robots.txt file [41], site owners can decide to allow or disallow their sites to be crawled. Even more significant is that changes to the robots.txt file are retroactive. This means that changes made at any time will affect whether the historical record held by the Wayback Machine is erased. The official statement on this policy includes the following key section:

While robots.txt has been adopted as the universal standard for robot exclusion, compliance with robots.txt is strictly voluntary. In fact most Web sites do not have a robots.txt file, and many Web crawlers are not programmed to obey the instructions anyway. However, Alexa Internet, the company that crawls the Web for the Internet Archive, does respect robots.txt instructions, and even does so retroactively. If a Web site owner decides he/she prefers not to have a Web crawler visiting his/her files and sets up robots.txt on the site, the Alexa crawlers will stop visiting those files and will make unavailable all files previously gathered from that site. [42]

As has been noted in several cases in discussions on archive.org [43] the retrospective application of changes to the robots.txt file threatens data permanence when it comes to domain squatting or simple changes of domain ownership and use. For example, recall that the founders of Kwetu.net lost their domain name by failing to renew it in time. If the current owner of Kwetu.net decided to make a small change to the current robots.txt file, all traces of Kwetu.net could vanish from the Wayback Machine forever. We address this later in the [conclusions](#).

Crawlers are similarly bound by legal constraints, such as in 2002, when the Wayback Machine removed entire domains from its archive to comply with a request from the Church of Scientology's lawyers (Miller, 2002). There is also a six-month embargo on new sites, thus letting much timely and fleeting content — of the kind at risk in Sri Lanka — slip away before it can be crawled unless it is selectively archived by a human.

Furthermore, and extremely relevant to the Gone Dark Project, it can be argued that “automated approaches to collecting Web data tend to stop short of incorporating the means to manage the risks of content loss to valuable Web documents” (Kenney, *et al.*, 2002). That is, it does not address the root causes. As our case studies show, just having a record of sites that used to be live is not a sufficient preservation strategy, although it is definitely an indispensable service to maintain.

Problems arise when existing copies of pages come to be seen as a permanent backup solution and when insufficient attention is paid to the content or pages that are allowed to disappear. Relying on archives run by national libraries around the world to do all the work — especially those collections that are not made public (BBC News, 2013) — can cause content creators as well as average Web users to become complacent and therefore not take proactive steps to save large amounts of valuable content in a more functional format. As a result, automated crawls might inadvertently diminish long-term health of Web resources by encouraging a passive approach to backups coupled with the misguided impression that nothing on the Internet can be lost forever, when in fact it happens all the time.

Conclusions

National libraries and digital archive organizations continue to draw attention to the dangers of disappearing Web content. They are setting standards and taking action to save entire Web sites or ephemeral social media content from being lost forever. Despite the many laudable successes of current digital preservation efforts, however, some weak spots remain as we have demonstrated. For instance, automated Web preservation is restricted to the indexable or “surface” Web. We are limited by an inability to foresee and therefore prevent content loss that falls outside of this. Many sites, including those in the case studies above, may hold large repositories of culturally significant information behind a pay wall, a registration system or in a database. Similarly, large collections of Web sites dispersed throughout an entire ecosystem — such as human rights and activist Web sites — are fragile especially because of the difficulty in tracking ownership or preserving whole sections of the Web that may become vulnerable all at once.

What vanishing content reveals is that there is a problem in self-organisation for the network of bodies that keep the Internet running [44]. There is a lack of clarity about who is responsible for archiving material so in some cases it falls between cracks and vanishes. Again we recognise that this is not a

new problem: it is at least as old as newspapers (no one was responsible for archiving a failing media business such as a local newspaper). So we may need more discussion of responsibility not only from Web archivists but also from content holders. Holders must be willing to have their content archived and made public (under certain conditions). And of course we need discussion of how all this is to be paid for. It is clear that archiving services are not without costs, especially as we think into the long term.

Sites that go dark do so for a variety of reasons from the financial to simple neglect and even malicious removal, but, as we have shown, that does not always mean that the original content has vanished. That we have been able, in the course of this project, to connect with relevant parties who report intentions to revive old pages given the right conditions means that better safeguards in place may be able to prevent such losses or hasten their return online. The following recommendations therefore consider what can be done to avoid future losses as well as suggesting ways to better preserve sites in imminent danger of vanishing.

A first step is to draw upon the excellent work already being done by Internet archivists to enhance the ease and regularity at which sites that fall within these grey areas can be saved. This means bridging the gap between professional librarians, academics, archivists and other dedicated individuals who can make worthwhile contributions. Paid, subscription services for Web site backup may be a good enterprise solution for the site owners with the means and access to make use of them (such as academic or corporate institutions), but the costs may be prohibitive to other users or they may simply lack of awareness of their existence. Encouraging dialogue between archive service providers and subject area experts will be the most effective way to save endangered or at-risk sites from going dark.

The background research we undertook revealed that many putative cases of disappearance were rather examples of link rot: the material was there but no longer at the same URL. This obscures a smaller number of actual instances of disappearance. Thus, arising from the case studies and interviews we have undertaken in this project, the broadest recommendation is to allow for more human intervention in the archival process such as appealing to subject experts who have first-hand knowledge of parts of the Web that are now at risk or may be in the future. Sanjana Hattotuwa exemplifies how specialist experience can inform better archiving practices based on actual needs and practices, while the Europa Film Treasures and Kwetu case studies show the importance of foresight and instilling good data management for long-term survival of Web content.

In addition, we have also encountered “inadvertent archivists”: these are mainly researchers or academics who have found that they have unintentionally become the curators of the only surviving copy of old Web content that they captured in the course of their research. Among these are some original Web site owners who may have old pages stored on their hard drives, but no means to restore them to the Web.

What can we do to help those who find themselves with knowledge, or in possession, of Web content, but who do not know what to do with it? Because of the complex reality of sites going dark, we find that combinations of human and technical solutions are necessary.

There are practical considerations to remedy vanishing Web sites which vary on a case-by-case basis. Depending on the type of site and/or repository of media or information in question, making the data public — or indeed accessing the content without the aid of insiders — can be difficult. For this reason, collaborative solutions are needed which bring together those content owners or researchers aware of imminent site losses with archival professionals who can assist them. Ideally, this would include

tailored services to better enable individuals in more perilous circumstances without the luxuries of institutional backing or secure funding sources to safeguard essential sites easily.

It is therefore important to continue developing tools to improve as well as open the archiving process to a wider audience. This will help to counteract the public perception that the entire Web is being backed up automatically when so much of it can remain at risk. Currently, the Internet Archive's Wayback Machine allows users to submit pages for archiving using the Save Page Now function, as do several other on-demand services. Yet none of these solutions reach out to the original site owner, make provisions for long-term preservation of original data, or endeavor to keep the site live. The backups they provide are also largely ineffective for recreating the site at a later date if the essential content is missing.

One technical solution to help bridge the gap may be an escrow-type backup system for the protection of endangered content. Such a solution would require archive professionals working closely with content owners or subject experts to produce preservation strategies that are easy to adopt, secure and flexible. The type of archive or backup system, its format and accessibility (open vs. restricted access) may vary depending on the needs of the organization or individuals wishing to secure their data and how sensitive that content may be. For instance, file format and integrity is a primary concern alongside legal requirements for preserving metadata to enable digital files to be used as a court record [45].

Working with experts would be of great value to help identify the type of sites we encountered in this project. At the same time, an interesting corollary is the need for improvements in the automated, technical side of Web preservation. Sanjana Hattotuwa adds this caveat: "machine and algorithmic curation can, with enough learning provided by analysing human curation, aid [the] archiving of content at risk esp. during violent conflict." [46] This can be invaluable in cases where the resources are simply not available to maintain fully staffed digital archives, such as in high-risk situations, with many non-profit companies, small organizations, poorer nations or NGOs. The same thing applies to small-scale sites whose owners are not available or otherwise up-to-date with good archiving practices.

Also worthy of consideration are open archive solutions to harness and analyze these aspects of human curation. Combining both technical and collaborative endeavors could result in a crowd-sourced solution that not only enabled users to submit sites at risk or already gone, but also then used submission data to predict other Web site candidates that may also be vulnerable. In the course of this project, we had expressions of interest from non-experts who wished to contribute more to efforts to save disappearing sites, but were unaware of any channels available to do so. Often, they could only offer an old URL for further investigation, which is where we were able to step in.

Lastly, developing solutions for safeguarding at-risk sites or reviving sites that have already gone dark requires improvements in how archives (and researchers) keep track of the disappearing Web over time. Inadvertently, this project has demonstrated the difficulties in identifying sites that may need help. It is certainly labor-intensive. One idea to remedy this lack of wider awareness about site losses is an early warning system for those parts of the Web that fall outside the scope of existing archival practices. An "endangered Web site alarm" could alert potential archivists of imminent content losses before or as they happen. For truly effective, proactive archiving solutions, this would go hand in hand with having clearer communication channels in place between archive service providers and others.

For example, while the Wayback Machine is an indispensable tool for Web research, as described above, several of its key restrictions limit its utility at present for pre-empting digital losses of sites

that are not easily crawled by Alexa. That said, the Internet Archive expressed willingness to allow access to its collections by “researchers, historians and scholars” [47], and in 2012, even experimentally offered researchers access to a full 80 terabytes of archived Web crawl data by request (Rossi, 2012). We believe that the data that the Internet Archive, Alexa, Google, other search engines and even Wikipedia collect may offer valuable insight into the evolution of the Web if researchers had access to certain information.

Rather than search manually for broken links to find URLs returning 404 errors as was done in the course of the Gone Dark Project, it would be much more useful if there were a system to export data from automated crawls that indicated persistent 404 errors within a given period of time to give researchers a chance to investigate further before the data is completely lost from the public Web. Similarly, logs of changes to robots.txt files (as noted above, these changes are retroactive and permanent and can wipe archive records) could alert researchers or Web preservationists of unforeseen losses as they happen. It might be that a change of robots.txt file which would trigger retrospective deletion could only go back as far as the current ownership of a domain. This is automatable so we recommend it to the Internet Archive. Another possible way of using 404 errors to promote archiving might be if they could delay the wiping of cached copies by Internet search services such as Google and Bing.


In addition, the Wayback Machine once had a functional search engine called Recall, designed by Anna Patterson [48]. Looking back on our research, it was difficult to locate important sites that have gone dark because it is nearly impossible to search historical Web content. Live search engines like Google cannot search defunct pages, while sites cannot be retrieved from most internet archives without the original URL. Enabling full-text searching of old pages would be ideal.

In all three case studies, a key lesson learned has been that a priority for improving Web preservation needs to begin at source, educating site owners and content producers so they understand the value of Web archiving. This is perhaps most key for high-risk sites or large repositories. But the education process needs to go both ways: the best practices for archiving are those which meet the current and future needs of those whose content would benefit from long-term storage and also those who will be able to make use of the content in future, whether to restore it to the public Web or to safeguard in a restricted archive.

Solutions to the problems of sites going dark will require more awareness from all parties involved. Making archiving initiatives more accessible, collaborative and lowering boundaries to participation (at present, interested parties must have “reasonably advanced programming skills” [49] to work with the Internet Archive’s data crawls, which is prohibitive for many) is a good start. Beyond simply collecting snapshots from old URLs, the long-term health of essential Web resources will depend on working with content owners to find permanent homes for at-risk data.

Recommendations summary

1. major service providers should consider maintaining backups as dark archives/escrow services
2. Internet archive services should provide a mechanism for “inadvertent archivists” to upload material (possibly not their own)

3. Internet archive services should provide a mechanism for experts to flag material as being at risk for urgent archiving [50]
4. Patterns in 404 errors should be investigated — can they predict data loss?
5. Google and Bing (etc.) consider responding to persistent 404 errors by passing cached copies to archive services.
6. Internet Archive respect changes of robots.txt file which would trigger retrospective deletion only as far as the current ownership of a domain. 

About the authors

Francine Barone is a social anthropologist and Internet researcher. Her ethnographic research focuses on the socio-cultural impacts of the digital age.

E-mail: fbarone [at] gmail [dot] com

David Zeitlyn is professor of social anthropology at the Institute of Social and Cultural Anthropology, University of Oxford. His field research is concentrated in Cameroon and he also works on archives and has been a pioneer of using the Internet to disseminate anthropology.

E-mail: david [dot] zeitlyn [at] anthro [dot] ox [dot] ac [dot] uk

Viktor Mayer-Schönberger is Professor of Internet Governance and Regulation at the Oxford Internet Institute, University of Oxford.

E-mail: viktor [dot] ms [at] oii [dot] ox [dot] ac [dot] uk

Notes

1. <http://www.archive.org>.

2. <http://archiveteam.org/index.php?title=Deathwatch>.

3. <http://www.webarchive.org.uk/>.

4. https://twitter.com/AT_Webarchive.

5. <http://www.archive.org>.

6. <http://www.digitalpreservation.gov>.

7. <http://www.internetmemory.org>.

8. Originally online at http://www.rsl.ox.ac.uk/isca/haddon/HADD_home.html.

9. See http://web.archive.org/web/20050415000000*/http://www.isca.ox.ac.uk/haddon/HADD_home.html. The 4 April 2005 is last working snapshot before they become 404 not found.

10. http://en.wikipedia.org/wiki/List_of_Google_products#Discontinued_products_and_services.

11. http://en.wikipedia.org/wiki/List_of_Yahoo!-owned_sites_and_services#Closed.2Fdefunct_services.
12. http://en.wikipedia.org/wiki/Windows_Live#Discontinued_services.
13. There are many others. Some of the most prominent are mentioned at https://en.wikipedia.org/wiki/Wayback_Machine.
14. Amad, 2010, p. 1; see also Zeitlyn (2012) and forthcoming.
15. The research for this section was undertaken by Nanjira Sambuli, iHub Research, Kenya.
16. See Stanford University's Library and Academic Information (Kenya) Resources listing: <http://www-sul.stanford.edu/depts/ssrg/africa/kenya.html>.
17. According to the company profile page accessible through the Wayback Machine: <http://web.archive.org/web/20030212235255/http://kwetu.net/about.asp>.
18. Source: <http://www.library.upenn.edu/news/86>.
19. Available via Wayback Machine: <http://web.archive.org/web/20060114003227/http://www.kwetu.net/partners.asp>.
20. <http://web.archive.org/web/20060114024930/http://www.kwetu.net/subscribers.asp>.
21. A cached copy of the front-end of the Kwetu.net search engine from 2003 is available from: <http://web.archive.org/web/20030812143045/http://kwetu.net/search.asp>.
22. http://www.openculture.com/2012/12/europa_film_treasures.html.
23. According to the site's original "About" page: http://web.archive.org/web/20130327054908/http://www.europafilmtreasures.eu/about_us.htm.
24. <http://www.lobsterfilms.com/ANG/index.php>.
25. Personal communication, 21 January 2015.
26. Personal communication, 18 September 2014.
27. <http://cinema.arte.tv/fr/magazine/europa-film-treasures>.
28. Personal communication, 21 January 2015.
29. <http://groundviews.org>.
30. <http://sitesatrisksl.wordpress.com>.
31. *Ibid.*

[32.](#) Personal communication, 3 May 2014.

[33.](#) Personal communication, 3 May 2014.

[34.](#) We note that this approach would not work for Web sites which access content via a database such as kwetu.net already discussed above.

[35.](#) <http://sitesatrisksl.wordpress.com/>.

[36.](#) The Human Rights Documentation Initiative at the University of Texas and Columbia University's Human Rights Web Archive, are both doing essential work for human rights Web preservation.

[37.](#) <https://www.archive-it.org/learn-more>.

[38.](#) <http://hoover.org>.

[39.](#) <https://archive-it.org/collections/4399>.

[40.](#) <https://archive-it.org/collections/2950>.

[41.](#) A file that contains requests from site owners that can prevent Web crawling software from crawling certain pages. See: <https://support.google.com/webmasters/answer/6062608?hl=en>. Note that not all Web crawlers respect robots.txt files. The Internet Archive does.

[42.](#) <http://archive.org/about/faqs.php#14>, (<http://perma.cc/528A-QMPH>, accessed 21 March 2015).

[43.](#) See <https://archive.org/post/406632/why-does-the-wayback-machine-pay-attention-to-robotstxt> (<http://perma.cc/NL3M-MNK9>) and <https://archive.org/post/188806/retroactive-robotstxt-and-domain-squatters> (<http://perma.cc/P6HL-VRWF>).

[44.](#) We are very grateful to *First Monday's* reviewers for suggesting that we acknowledge this point explicitly — and for other points made in the review.

[45.](#) Sanjana Hattotuwa, personal communication, 11 November 2014.

[46.](#) Personal communication, 11 November 2014.

[47.](#) <http://web.archive.org/web/20090924112618/> and http://www.archive.org/web/researcher/intended_users.php.

[48.](#) <http://web.archive.org/web/20031204221423/ia00406.archive.org/about.html>.

[49.](#) Explained

here: http://web.archive.org/web/20090924112618/http://www.archive.org/web/researcher/intended_users.php.

[50.](#) Manual archiving is possible using services such as “Save Page Now” and “Archive-It”. However, these share the same problem as the crawler-based automatic services of not having access to the content of Web-searchable databases.

References

Note: We have created perma.cc archive copies for our online sources. For completeness, we give both URLs although the permac.cc URL passes through to the original URL if it is still available, serving the archived copy only if the original URL generates a 404 error.

P. Amad, 2010. *Counter-archive: Film, the everyday, and Albert Kahn's Archives de la Planète*. New York: Columbia University Press.

Archive-It, 2014. "Only 41% of Occupy Movement URLs accessible on live Web," *Archive-It Blog*, at <https://archive-it.org/blog/only-41-of-occupy-movement-urls-accessible-on-live-web>, accessed 13 November 2014; <http://perma.cc/RJF6-CRXL>, accessed 24 April 2015.

J. Barksdale and F. Berman, 2007. "Saving our digital heritage," *Washington Post* (16 May), at <http://www.washingtonpost.com/wp-dyn/content/article/2007/05/15/AR2007051501873.html>, accessed 14 February 2014; <http://perma.cc/FNX5-Y97X>, accessed 24 April 2015.

BBC News, 2013. "Web archive goes live but not online," *BBC News* (19 December), at <http://www.bbc.com/news/technology-25446913>, accessed 13 November 2014; <http://perma.cc/5FGP-BCHQ>, accessed 24 April 2015.

BBC News, 2010. "British Library warns UK's Web heritage 'could be lost'," *BBC News* (25 February), at <http://news.bbc.co.uk/2/hi/technology/8535384.stm>, accessed 11 November 2014; <http://perma.cc/U538-5F5M>, accessed 24 April 2015.

O. Bhuiyan, 2014. "Technorati — the world's largest blog directory — is gone." *Business 2 Community* (16 June), at <http://www.business2community.com/social-media/technorati-worlds-largest-blog-directory-gone-0915716>, accessed 11 November 2014; <http://perma.cc/4NZC-GQM4>, accessed 24 April 2015.

C. Dewey, 2014. "How Web archivists and other digital sleuths are unraveling the mystery of MH17," *Washington Post* (21 July) (available on-line: <http://www.washingtonpost.com/news/the-intersect/wp/2014/07/21/how-web-archivists-and-other-digital-sleuths-are-unraveling-the-mystery-of-mh17/>), accessed 13 November 2014; <http://perma.cc/7AMQ-K8ZP>, accessed 24 April 2015.

Economist, 2000. "The hopeless continent," *Economist* (13 May), at <http://www.economist.com/node/333429>, accessed 12 November 2014; <http://perma.cc/E6L8-Q3UT>, accessed 24 April 2015.

S. Hattotuwa, 2008. "Websites at risk — Archiving information on human rights, governance and peace," *ICT for Peacebuilding (ICT4Peace)*, at <http://ict4peace.wordpress.com/2008/04/02/websites-at-risk-archiving-information-on-human-rights-governance-and-peace/>, accessed 13 November 2014; <http://perma.cc/6EFV-X25D>, accessed 24 April 2015.

Hoover Institution, 2014. "Archivists capture evidence in Malaysia Airlines Flight 17 crash" (25 July), at <http://www.hoover.org/news/archivists-capture-evidence-malaysia-airlines-flight-17-crash>, accessed 13 November 2014; <http://perma.cc/54D2-RR6B>, accessed 24 April 2015.

- Internet Archive, 2014. "Wayback Machine hits 400,000,000,000!" *Internet Archive Blogs* (9 May), at <http://blog.archive.org/2014/05/09/wayback-machine-hits-400000000000/>, accessed 11 November 2014; <http://perma.cc/RW5Y-2PSQ>, accessed 24 April 2015.
- A. Kenney, N. McGovern, P. Botticelli, R. Entlich, C. Lagoze and S. Payette, 2002. "Preservation risk management for Web resources: Virtual remote control in Cornell's Project Prism," *D-Lib Magazine*, volume 8, number 1, at <http://www.dlib.org/dlib/january02/kenney/01kenney.html>, accessed 13 November 2014; <http://perma.cc/BLQ8-D4Z2>, accessed 24 April 2015.
- S. Lohr, 2010. "Library of Congress will save tweets," *New York Times* (14 April), at <http://www.nytimes.com/2010/04/15/technology/15twitter.html>, accessed 11 November 2014; <http://perma.cc/QA6Y-GU69>, accessed 24 April 2015.
- V. Mayer-Schönberger, 2009. *Delete: The virtue of forgetting in the digital age*. Princeton, N.J.: Princeton University Press.
- E. Miller, 2014. "Sherman, set the Wayback Machine for scientology," *LawMeme* (24 September), at <http://web.archive.org/web/20141025203224/http://lawmeme.research.yale.edu/modules.php?name=News&file=article&sid=350>, accessed 13 November 2014; <http://perma.cc/2JRV-5CX7>, accessed 24 April 2015.
- D. Quenqua, 2009. "Blogs falling in an empty forest," *New York Times* (5 June), at <http://www.nytimes.com/2009/06/07/fashion/07blogs.html>, accessed 11 November 2014; <http://perma.cc/Z5F6-FRGJ>, accessed 24 April 2015.
- J. Rosen, 2010. "The Web means the end of forgetting," *New York Times* (21 July), at <http://www.nytimes.com/2010/07/25/magazine/25privacy-t2.html>, accessed 13 November 2014.
- A. Rossi, 2012. "80 terabytes of archived Web crawl data available for research," *Internet Archive Blogs* (26 October), at <http://blog.archive.org/2012/10/26/80-terabytes-of-archived-web-crawl-data-available-for-research/>, accessed 22 December 2014; <http://perma.cc/UPF4-3MQ4>, accessed 24 April 2015.
- H. SalahEldeen and M. Nelson, 2012. "Losing my revolution: How many resources shared on social media have been lost?" *arXiv* (13 September), at <http://arxiv.org/abs/1209.3026>, accessed 11 November 2014; <http://perma.cc/U3Q2-R8YM>, accessed 24 April 2015; also in: P. Zaphiris, G. Buchanan, E. Rasmussen and F. Loizides (editors). *Theory and practice of digital libraries: Second international conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings. Lecture Notes in Computer Science*, volume 7489. Berlin: Springer-Verlag, pp. 125–137. doi: http://dx.doi.org/10.1007/978-3-642-33290-6_14, accessed 24 April 2015.
- N. Taylor, 2014. "The MH17 crash and selective Web archiving," *The Signal: Digital Preservation* (28 July), at <http://blogs.loc.gov/digitalpreservation/2014/07/21503/>, accessed 13 November 2014; <http://perma.cc/7TGU-BBWJ>, accessed 24 April 2015.
- N. Taylor, 2011. "The average lifespan of a Webpage," *The Signal: Digital Preservation* (8 November), at <http://blogs.loc.gov/digitalpreservation/2011/11/the-average-lifespan-of-a-webpage/>, accessed 11 November 2014; <http://perma.cc/FL5X-R285>, accessed 24 April 2015.

Z. Whittaker, 2010. "Facebook does not erase user-deleted content," *ZDNet* (28 April) at <http://www.zdnet.com/blog/igeneration/facebook-does-not-erase-user-deleted-content/4808>, accessed 13 November 2014; <http://perma.cc/2ECY-HTRQ>, accessed 24 April 2015.

A. Wright, 2009. "Exploring a 'deep Web' that Google can't grasp," *New York Times* (22 February), at <http://www.nytimes.com/2009/02/23/technology/internet/23search.html>, accessed 13 November 2014; <http://perma.cc/MG5L-QRBW>, accessed 24 April 2015.

D. Zeitlyn, forthcoming. "Looking forward, looking back," *History and Anthropology*.

D. Zeitlyn, 2012. "Anthropology in and of the archives: Possible futures and contingent pasts. Archives as anthropological surrogates," *Annual Review of Anthropology*, volume 41, pp. 461–480. doi: <http://dx.doi.org/10.1146/annurev-anthro-092611-145721>, accessed 24 April 2015.

J. Zittrain, K. Albert and L. Lessig, 2013. "Perma: Scoping and addressing the problem of link and reference rot in legal citations," *Harvard Public Law Working Paper*, number 13–42, at <http://papers.ssrn.com/abstract=2329161>, accessed 24 February 2014; <http://perma.cc/4DVN-DYS8>, accessed 24 April 2015; also in *Harvard Law Review*, volume 127, number 4 (2014), at <http://harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations/>, accessed 24 April 2015.

Editorial history

Received 28 January 2015; revised 30 March 2015; accepted 7 April 2015.

Copyright © 2015, *First Monday*.

Copyright © 2015, Francine Barone, David Zeitlyn, and Viktor Mayer-Schönberger.

Learning from failure: The case of the disappearing Web site
by Francine Barone, David Zeitlyn, and Viktor Mayer-Schönberger.

First Monday, Volume 20, Number 5 - 4 May 2015

<https://firstmonday.org/ojs/index.php/fm/article/download/5852/4456>

doi: <http://dx.doi.org/10.5210/fm.v20i5.5852>

What Do you Mean by Archive? Genres of Usage for Digital Preservers

February 27, 2014 by [Trevor Owens](#)

One of the tricks to working in an interdisciplinary field like digital preservation is that all too often we can be using the same terms but not actually talking about the same things. In my opinion, the most fraught term in digital preservation discussions is “archive.” At this point, it has come to mean a lot of different things in different contexts. It can mean so many different things that [some in digital preservation are reluctant to use the term writ large](#) (<http://www.avpreserve.com/blog/why-i-wont-be-using-the-word-archive-anymore/>) . I wanted to spend a few moments putting text on a URL that anyone can reference from here on out when they need to try and parse and disambiguate what we mean by archive. For a some related reading, I'd suggest checking out Kate Theimer's [Archives in Context and as Context](#) (<http://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/>) and the role of “the professional discipline” in archives and digital archives. (<http://www.archivesnext.com/?p=3683>)

I'd stress here that I'm not really interested in telling people what is and isn't an archive. Instead, I'm interested in 1) helping people ensure that they aren't talking past each other and 2) briefly starting to suss out the resonances between these different usages. I would love to hear more perspectives on usage of the term and resonances between those uses in the comments. In many different contexts the term archive carries with it significant weight, the term often brings with it notions of longevity, safe keeping, order and concerns with authenticity, it's about items or records that hang together for good reason. To varying extents, across each of the uses I articulate here I think we see these points surface. My objective here is not to exhaustively describe any of these ways the word is used, but just too briefly gesture toward different usages. I should stress that this is how I sort out some of the different usages of the terms. I invite readers to suggest additional and or different usages and comment these below the post.

Archive as in Records Management

In an organizational context, [an archives](#) (<http://www2.archivists.org/glossary/terms/a/archives>) is often the place in the organization that is required to retain and organize records of the organization. So a radio station, or a hospital, or a financial services company needs to keep around copies of records of its operation for a range of reasons (litigation, tax purposes, posterity, compliance with regulations, etc.). In this case, the archive serves the purpose of organizing, maintaining records and materials for use by the organization. In this case, a big part of the work of an archive is to make sure they are keeping around only what is deemed to be useful for particular future use cases.

Archive as in “The Papers of So and So”

One of the specific senses that archivists will use the term archives is to describe a particular kind of collection. Effectively, an archive is a kind of collection of materials that hang together for a very particular reason. An archive is either the papers of some particular person or the papers or records of a particular organization. What makes it an archive is the fact that the items and records in the collection represent “[fonds](#) (<http://www2.archivists.org/glossary/terms/f/fonds>) ” a particular name for a collection that are the result of the ongoing work of the individual or organization. The words “natural” and “organic” generally come into play here, the idea being that the archive is a collection of items and records that exist as a whole. To contrast with this,

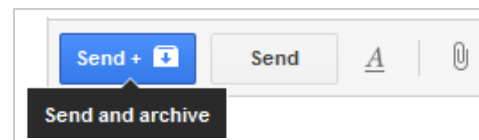


([//blogs.loc.gov/digitalpreservation/files/2014/02/Manuscript-Division-stacks-with-acid-free-containers](http://blogs.loc.gov/digitalpreservation/files/2014/02/Manuscript-Division-stacks-with-acid-free-containers))
[//lcweb2.loc.gov/ammem/mchtml/speci](http://lcweb2.loc.gov/ammem/mchtml/speci)
 . Manuscript Division Slide Collection

an archivist might refer to a collection of rare books pulled together by a collector over time an “artificial” collection. Artificial in this case is not to say that it’s “bad” just that the collection was assembled as a set of materials after the fact.

Archive as in “Right Click -> add to Archive”

For most people, the most common usage of the term archive is likely from a context menu in computing. In many operating systems you can simply right click on some icon for a file and click “add to archive” or “create archive.” In these cases, borrowing on a legacy of usage of the term more generally in computing, this ends up meaning stick it into some kind of compressed container file. In this vein, the term archive is largely tied to the idea of “back-up.” Effectively, the archived copy of these files is slightly more difficult to get to but right at your fingertips nonetheless.



(<https://blogs.loc.gov/digitalpreservation/files/2014/02/right-click-send-to-archive.png>)

Example of archive used in web mail.

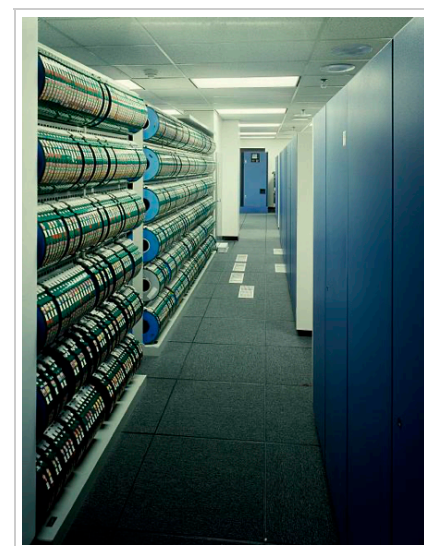
Usage of the term in web applications, like web email clients, is very similar. In the case of many web mail systems the archive is simply all of your emails that you haven’t deleted and are not in your inbox. In the logic of “piling vs. filing” (<http://lifesacker.com/238339/file-organization-strategies--filing-versus-piling>) this actually makes sense. In the past, you might have organized your correspondence and bills in a particular and structured fashion, keeping only what you needed for the future and deliberately putting it where it would be easy to find in the future. That filing process for managing records is much more inline with what archivists mean by archive. As email has shifted further and further toward something that people expect to be able to simply do full text search against the term archive has come along with it, but the fact that folks now generally just let it pile up in one big thing called “archive” that they search against is very different from the deliberate organized thing that archivists are generally talking about.

Archive as in “Tape Archive”

When IT people use the term archive they are generally talking about a piece of hardware. At the start of each of the [Library of Congress storage architecture meetings](http://www.digitalpreservation.gov/meetings/storage13.html) (<http://www.digitalpreservation.gov/meetings/storage13.html>) we generally need to begin with this vocabulary discussion. As an example, many large organizations use a [HSM](http://en.wikipedia.org/wiki/Hierarchical_storage_management) (http://en.wikipedia.org/wiki/Hierarchical_storage_management) , a [hierarchical storage management system](http://en.wikipedia.org/wiki/Hierarchical_storage_management) (http://en.wikipedia.org/wiki/Hierarchical_storage_management) , that maintains different tiers of storage that have distinct performance requirements. At this point, the top level might be a relatively small amount of expensive but fast flash memory, below that might be a larger pool of spinning disk storage, below that you would likely find something called the “archive” layer. In this case, archive means tape archive. Magnetic tape remains the cheapest medium (you can store a lot more data on tape for a lower cost than disc) but it is significantly less responsive. So it is going to take you time to get the information back from tape. So within the design of a storage system, the stuff you need to keep around but don’t need to access that often, or your back up copies etc. ends up on the biggest but cheapest tier of your storage system.

The definition here relies on a long history of using the term archive as a synonym for magnetic tape storage systems. The [file format .tar](http://en.wikipedia.org/wiki/Tar_(computing)) ([http://en.wikipedia.org/wiki/Tar_\(computing\)](http://en.wikipedia.org/wiki/Tar_(computing))) , a way to package data for storage, itself stands for “tape archive.” This use of the term archive goes back to 1940s computer systems architecture. In the original context it referenced online vs. offline storage. The reels of tape were quite literally “off line,” the reel had to be located and mounted before data became accessible in contrast to things like a magnetic core at the time, and later random access memory.

Archive in “Web Archive”



Computer data storage in a modern office building, taken during the 1980s (<http://www.loc.gov/pictures/item/2011634402/>) , Photographs in the Carol M. Highsmith Archive, Library of Congress, Prints and Photographs Division.

Many organizations are now in the business of harvesting content from the web for long term access and preservation. In these cases, tools like [Heritrix](#), an open source web webcrawler (<https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>), are sent out to grab all of the rendered content of a webpage they can get ahold of and, within defined parameters, the other pages that link to it and all their associated files. As part of this collection process, the tools log information about the date and time that the data was collected. At this point, tools store that content in [WARC](#) (<http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml>) files, or Web Archive files, which can then be played back via tools like the [Wayback machine](#) (http://en.wikipedia.org/wiki/Wayback_Machine). So there is a lot of information in here that can be used to assert the authenticity of the data, how a particular URL presented itself to Heritrix and how Heritrix interpreted it at a particular moment in time. With that said, it's much more in keeping with the computing usage of archive as a back-up copy of information than the disciplinary perspective of archives.

Archive as in “Digital Archive”

At this point, there are a lot of digital collections that are using the term archive that don't necessarily square with how archivists have been using the term. For instance, the [September 11th Digital Archive](#) (<http://911digitalarchive.org/>), the [Bracero Archive](#) (<http://braceroarchive.org/about>) the [The Shelley-Godwin Archive](#) (<http://shelleygodwinarchive.org/>) are good exemplars of some of the diversity of this usage. In each case, an effort was undertaken to bring collect or bring together related materials. The September 11th digital archive is a crowdsourced collection of materials related to the attacks, the Bracero Archive is a digitized collection of oral history interviews with individuals involved in the Bracero guest worker program, and The Shelley-Godwin Archive brings together digitized copies of primary manuscript sources related to a particular family. The origin of this usage is anchored in Jerome McGann's work on the [Rossetti Archive](#) (<http://www.rossettiarchive.org/>), which McGann had developed grounded in a theoretical perspective of the [potential that hypermedia brought to allow for the creation of new kinds of archives](#) (<http://www2.iath.virginia.edu/public/jjm2f/rationale.html>). Alongside this usage, digital archive has also been used as a term to refer to born digital materials processed as part of a more traditional notion of an archive. In this case, see usage of [“the born digital archives of Salman Rushdie](#) (http://www.youtube.com/watch?v=oiqHv_SofNo)”.



(<https://blogs.loc.gov/digitalpreservation/files/2014/02/wlaw-blog.jpg>)

[Wendy's Blog: Legal Tags](#) (<http://www.loc.gov/item/lcwa00090233>), Legal Blawgs Web Archive, Law Library of Congress

Some archives purists might call all of these [“artificial” collections](#) (<http://www2.archivists.org/glossary/terms/a/artificial-collection>). I however wouldn't. I don't think this is so much about the computing terminology invading the space, but instead another tradition in which systematically collected materials have been called archives within cultural heritage organizations. Folklife archives, for example the [American Folklife Center Archive](#) (<http://www.loc.gov/folklife/archive.html>), at the Library of Congress, have long worked to acquire [ethnographic field collection](#) (<http://www.loc.gov/folklife/ethno.html>)'s for the archive. In these cases, folklorists have gone out and made field recordings and then worked with archivists to organized them for access. With this said, its valuable to recognize that generally the term digital archive carries this language and meaning as opposed to the canonical repository for the “papers of so and so” or the records management terminology. That is, digital archives hang together as [“a conscious weaving together of different representational media](#) (<http://www.loc.gov/folklife/ethno.html>)”.

For another take on the idea of digital archives see Kate Theimer's recent presentation at the American Historical Association's annual meeting, [A Distinction worth Exploring: “Archives” and “Digital Historical Representations.”](#) (<http://www.archivesnext.com/?p=3645>)

Notions and Considerations of “The Archive”

The last category I am including here is about theorizing “the Archive.” A broad range of work in literary and media theory focuses attention on “the Archive.” Here I am thinking of Foucault's notion of “the Archive” in *The Archeology of Knowledge*, Derrida's perspective in *Archive Fever*, and Kittler and Wolfgang Ernst's notions of archives in [Media Archeology](#) (<https://blogs.loc.gov/digitalpreservation/2012/10/media-archaeology-and-digital-stewardship-an-interview-with-lori-emerson/>). For the most part, this body of work is less about what goes on in an individual archives and is more about the role of “the archive” in society writ large or the idea of “the archive” as traces of the past in objects. For example, for Foucault, “the Archive” is not so much an individual set of materials but a term for the entirety of historical records/evidence that exists to work from. These theoretical takes on “the

archive” can be frustrating to many archivists, as much of this work does not engaged with the professional practices of archives or with “archival theory (<http://readingarchives.blogspot.com/2009/03/archival-theory.html>) ,” the body of scholarship which archivists themselves have been building through ongoing practice and research since at least the French revolution.

At the institutional level, discussions of “the archive” are broadly useful for reflecting on the social roles that archives play in culture. Further, a considerable amount of this work in the Media Archeology and Media Theory traditions focus on processes of inscription and embedded logic of different media ([optical media \(http://www.electronicbookreview.com/thread/criticalecologies/opticalogic\)](http://www.electronicbookreview.com/thread/criticalecologies/opticalogic) , [gramaphones \(http://hydra.humanities.uci.edu/kittler/intro.html\)](http://hydra.humanities.uci.edu/kittler/intro.html) , [databases \(http://switch.sjsu.edu/web/v5n3/J-1.html\)](http://switch.sjsu.edu/web/v5n3/J-1.html) , [the MP3 format \(http://soundstudiesblog.com/2012/11/05/review-jonathan-sterne-mp3-the-meaning-of-a-format/\)](http://soundstudiesblog.com/2012/11/05/review-jonathan-sterne-mp3-the-meaning-of-a-format/) , etc) which are increasingly important genres of artifacts and records that archives are themselves tasked accessioning. Kirshenbaum’s *Mechanisms: New Media and the Forensic Imagination* (<http://digitalhumanities.org/dhq/vol/3/2/000048/000048.html>) is itself an invaluable exemplar of how work from these media theory traditions can combine with archival theory to produce scholarship that [directly informs \(http://www.ils.unc.edu/caltee/p57-woods.pdf\)](http://www.ils.unc.edu/caltee/p57-woods.pdf) the [development of tools and practices for practicing archivists \(http://www.bitcurator.net/\)](http://www.bitcurator.net/) . Again, these [broad and interdisciplinary conversation about archives \(https://dspace.mit.edu/handle/1721.1/35687\)](https://dspace.mit.edu/handle/1721.1/35687) can be quite useful to both those working in and outside archives (<http://wsampson.wordpress.com/2011/04/10/from-my-archives-derridas-archive-fever/>) .

So, are there other definitions I’m missing? Have I got any of the lineage wrong on this? I’d love to continue this discussion in the comments.

Thanks to Matthew Kirshenbaum, Nicki Saylor, and Kate Theimer for comments and suggestions for improvements to this post.

Posted in: [Digital Content, Education and Training](#)

[17 Comments](#) | [Add a Comment »](#)

17 Comments

1. **Dean C. Rowan**

February 27, 2014 at 12:05 pm

It’s always worth reviewing OED’s take on a term like “archive,” including its etymology. There is a distinct sense of the official to the term, as ἀρχή means “government.” Clearly, the several meanings you have set out above are abstracting from that sense, using it figuratively, which is how a lexicon evolves. I’d argue that the theorized Archive goes a step further. Foucault, Derrida, et al. want to impose on the term a sense of sinister agency, of an official mechanism of control. In this respect, the Archive is the flip-side of the Library, which is often rendered in spiritual terms. Libraries are “sacred,” “hushed,” transcendent, quiet places for contemplation, etc. I’m not especially fond of these murky ascriptions.

Another work worth examining is Cornelia Vismann’s *Files* , which focuses on the contents of archives in the records management sense.

2. **Bill LeFurgy**

February 27, 2014 at 12:19 pm

This is a good overview, and illustrates that “archive” and it’s pluralized variant have irrevocably escaped the semantic confines of their specialized origins.

Some of us do remember a time when the terms were used with more precision, and still might quibble (in total futility, no doubt) with how they are applied today. I personally see “records management” as something quite differ than “archives,” most especially in the context mentioned above. Organizations have records management programs to “serve the purpose of organizing, maintaining records and materials for use by the organization.” Organizations have archival programs if they keep some of those records for their enduring (i.e. permanent) value. The key difference is that records management is focused on disposition: keeping records for some length of time (often quite briefly) and then disposing of them. Records management and archives are clearly linked, but they are not the same, to my mind.

Now, in possible contradiction of myself, let me also say that I endorse the more recent usage of the term “personal archiving” and “personal archives” in reference to the material, often digital, that individuals create and maintain about themselves and their families. Some of this material might not be kept permanently (whether accidentally or by design) but people are now in possession of sizable personal collections that have important current and future personal value. Awareness is starting to dawn on many about this value, and that they probably, at some point, need to do something with their material. As you say, “archive” conveys notions of “longevity, safe keeping and order,” and I think these are the right concepts that people need to consider in connection with their personal digital material.

Now, actually applying those concepts is a tricky business, but “personal archiving” provides the right motivation. Besides, I’m not sure there is a better term. Something like “personal information management” might be more descriptive, but it’s a snooze in terms of impact.

3. **Carl Fleischhauer**

February 27, 2014 at 12:45 pm

Thanks as always for a helpful exploration of terms and usage. Your report also reminds us of the limits to surgical precision in speaking and writing — even when we wish to be. In your case, this is demonstrated by your apt use of `_ostensive_` rather than “dictionary” definitions for the term `_archive_` in its many contexts.

As I read, my thoughts drifted off into one of the next layers (beyond your immediate topic), puzzling over how we find meaning in the contents of an, um, archive(s). And this made me recall the wonderful insider term `_diplomats_`. As the redoubtable Wikipedia tells us: “. . . a scholarly discipline centred on the critical analysis of documents – particularly, but not exclusively, historical documents . . . focuses on the conventions, protocols and formulae that have been used by document creators, and uses these to increase understanding of the processes of document creation, of information transmission, and of the relationships between the facts which the documents purport to record and reality.” (<http://en.wikipedia.org/wiki/Diplomatics>). If we are to study the entities found in, well, a set or batch of entities (“archive”), and if we wish to get at the relationships between the stated facts and reality (sometimes we don’t!), we will depend upon explanations of what sort of archive this is, and upon its organization. Those of us who work in libraries and (yes) archives are supremely aware of this need and this naturally prejudices us toward certain uses of the term.

4. **Web Webster**

February 27, 2014 at 6:08 pm

The word “GENRE”

Oh, but I see how this will do wonders for precision and recall, how it will enhance rankings for the word “archival”, and certainly provide more false drops and eventually, lead to calls to the support desk of asking, “Why am I getting the wrong information?”

Not so much a misnomer, but more of how a word such as GENRE once meant “mongrel”, then meant offspring of a tame sow and wild boar, then child of a freeman and slave, then something such as cross-breeding—to a contemporary gas electric combo Prius.

It will do wonders for voice input with a string of NOT s

Web

5. **Stevan Lockhart**

February 28, 2014 at 6:12 am

It was ever thus. The term “database” struggles similarly. Some people talk of a database of information which others would describe as a dataset. Some refer to the underlying technical system, others to a data management application and so on. In the digital era, the term archive is similarly conflated to some as a web application, not the process of selection, storage and curation that may be implied.

The additional difficulty pointed out at the beginning of the piece is differencing expectations of terminology by practitioners and, for want of a better term, their non-specialist management who may promote the significance of an evident outcome while misunderstanding the role of design and process underlying that outcome. In this sense, trust among the parties communicating is especially important.

6. Eldin Rammell

[February 28, 2014 at 11:04 am](#)

There is also a nuance on “Archive as in Records Management” that is not covered in your otherwise excellent summary. In many organizations there exists “records centers” and “archives”. Your description of archives as “the place in the organization that is required to retain and organize records of the organization” could equally apply to both. The distinguishing factors are that (a) archives are generally for records that are inactive and (b) archives are under the control of an archivist. A record or collection of records are typically organized in records centers, perhaps under special security conditions and environmental controls, but these are often not considered archives if the records are active or semi-active. Once a matter is closed and the records are transferred to “deep storage”, this is “the archive”. On my second point, archives in commercial organizations often have an individual identified as “the archivist”.... this is often even a regulatory requirement (e.g. OECD Principles of Good Laboratory Practice). The identification of an archivist role would thus also be helpful in distinguishing archives from records centers.

7. Jason Cooper

[February 28, 2014 at 12:56 pm](#)

Archives and records management are clearly related, but not in all contexts. They are particularly intertwined in the corporate and governmental worlds, when materials move from a state of active use, through a period of less active use. At some point, some records are recognized to have enduring value beyond the intentions of their original creation. The records manager transfers these records to the archives (though in some cases these people are in fact one person) where they are properly described, filed, and preserved. This is how the records of the US Government are treated.

From your examples, records management deals with those records that exist for litigation, tax purposes, and compliance with regulations, while an archives holds them for posterity after the business needs have been met. Interestingly, the use of archive in the email example follows a similar vein of thinking – messages go from “I’m using that” to “This has some value, so I’m going to keep it somewhere else.”

8. Maarja Krusten

[February 28, 2014 at 4:30 pm](#)

Excellent summary! Have bookmarked it for future reference. Thanks also for the reference to Kate Theimer, guru to many archivists.

One small piece of supplemental information on the side. Within the federal government, the venue with which I am most familiar, some records with value for “posterity” are used not only by the “business owners” but by federal historians doing research to support employees of all ranks in an agency or department. Research may provide information for testimony statements, policy making, etc.

Such people are knowledge accountable officers. In agencies that have no historian, a records officer sometimes handles some of the records-search function, without the full scope of historian duties.

The records may be held by the agency mission or mission support units or elsewhere in the agency (such as a records room or library annex) or at a Federal records center. In theory, some of the records that a federal historian and business owners use (for different purposes) while active at some later point change in designated status (at least) to “inactive.” Such permanently valuable records then are become a part of the collections held for the American people in the U.S. National Archives.

At that point, the change in legal title affects the means of external access. It changes from the agency Freedom of Information Act request handling process to disclosure determination by the National Archives.

9. John Rees

[February 28, 2014 at 4:57 pm](#)

In terms of archival descriptive practices, and theory as I was taught in Archives 101 long ago, your phrase “‘artificial’ collection” is redundant.

In archival practice a ‘collection’ is naturally ‘artificial’ just as you describe, e.g. the “John Doe Collection of Louisiana Farm Workers Ethnographic Field Recordings.” One should not use the phrase ‘the collection of John Doe Papers’ or “John Doe’s collection over there at the university.”

But perhaps mine is an artifact of one person's teaching. SAA's Glossary of Archival Terminology conflates the phrase as well: <http://www2.archivists.org/glossary/terms/c/collection>

DACS rule 2.2.18 similarly distinguishes between describing 'creators' and 'collectors' but does not delve too far into the semantics or metaphysics of the terms.

10. **Helen Halmay**

February 28, 2014 at 5:10 pm

This is a very interesting, important topic. If you-all ever agree on a definition of "archive," I'd like to hear about it, and publish it in my newsletter (with your permission and attribution to you and the Library of Congress, of course). NOTE: you wrote: "...to try and parse and disambiguate what we mean by archive." It should be: "...to try to parse and ... etc." Just keeping you on your toes. Helen Halmay, Editor – Adelante, member newsletter for the Congress of History of San Diego and Imperial Counties, California

11. **Greg Bak**

March 3, 2014 at 8:35 am

Great post! It is nice to have this all pulled together.

Here is another one for you, from OAIS.

Archive: An organization that intends to preserve information for access and use by a Designated Community.

This definition is consistent with the rest of OAIS: focussed on access and use, always relative to the needs of a designated community, and emphasizing the social and operational dynamics of the archival organization rather than the technology used for preservation or delivery.

12. **Michael Winter**

March 13, 2014 at 7:55 pm

It only adds to the confusion that this blog post and the series of replies it occasioned so well clarifies, that "archive(s)" for a very long time has been used as part of the titles of a significant number of scholarly journals, e.g. Archives of Psychiatry and Psychotherapy, Archives of Public Health, Archives of Sexual Behavior, and many others. The examples listed here, by the way, are all for current periodicals still using these titles, that are in no way archives in any sense that most of us would recognize.

13. **Christie Peterson**

March 19, 2014 at 1:41 pm

Some of my colleagues here at Johns Hopkins have adopted the term "archive" to refer to the layer in a data management stack that manages fixity and integrity. See <https://www.youtube.com/watch?v=F6iYXNvCRO4&feature=plcp> for a full explanation. While this use of "archive" may raise my hackles a bit as an archivist, it makes for very useful shorthand during discussions about digital preservation.

14. **Katherine D. Harris**

September 17, 2014 at 1:35 pm

In less detail (due to word count constraints), I've rehearsed some of these debates for my entry in the Johns Hopkins Guide to Digital Media (JHUP 2014). There was a huge kerfuffle about archive being used in literary studies and the imposition of "database" in 2007-2009 (PMLA, Digital Humanities Quarterly). And, there have been quite a few disagreements among professionals about the use/abuse/colonization of "archive" by literary studies and Digital Humanists.

In an attempt to avoid replicating my entry for the JHGDM, I've crafted a further response to this idea of the archive for another project, but the keyword entry really is an extension of my thoughts from the JHG. (Due to copyright restrictions, I can't post the JHG entry online, though.) I welcome comments on that rough draft of "archive": <http://culturedigitally.org/2014/09/archive-draft-digitalkeywords/>

15. **Irene M.**

October 15, 2014 at 3:50 am

Very useful information. Just a thought, would a digital repository like an Institutional Repository in any way be regarded as an archive of sorts? In the sense that it acts as long term access point for information?

16. **Christopher Mule**

November 11, 2014 at 1:52 pm

Wow! This was incredibly helpful. Much appreciated.

17. **Julia Alaniz**

April 22, 2018 at 3:34 pm

When my father purchased land acreage, he would take documents to get recorded at the county courthouse.

Once the documents were formally recorded, he referred to them as “archivos”—archives (noun) because they were already “archivados”—archived (participle adjective modifier): filed, cataloged, registered, stored, etc.; the documents were archived (verb) in the history of the property.

Hmm...confusing...

Disclaimer

This blog does not represent official Library of Congress communications.

Links to external Internet sites on Library of Congress Web pages do not constitute the Library's endorsement of the content of their Web sites or of their policies or products. Please read our [Standard Disclaimer](#).