The Atlantic

Raiders of the Lost Web

If a Pulitzer-finalist 34-part series of investigative journalism can vanish from the web, anything can.

Maison Bonfils / Library of Congress / Zak Bickel / The Atlantic

ADRIENNE LAFRANCE

OCT 14, 2015 | TECHNOLOGY

The web, as it appears at any one moment, is a phantasmagoria. It's not a place in any reliable sense of the word. It is not a repository. It is not a library. It is a constantly changing patchwork of perpetual nowness.

You can't count on the web, okay? It's unstable. You have to know this.

Digital information itself has all kinds of advantages. It can be read by machines, sorted and analyzed in massive quantities, and disseminated instantaneously. "Except when it goes, it really goes," said Jason Scott, an archivist and historian for the Internet Archive. "It's *gone* gone. A piece of paper can burn and you can still kind of get something from it. With a hard drive or a URL, when it's gone, there is just zero recourse."

There are exceptions. The Internet Archive's Wayback Machine has a trove of cached web pages going back to 1996. Scott and his colleagues are saving tens of petabytes of data, chasing an ideal that doubles as their motto: Universal Access to All Knowledge. The trove they've built is extraordinary, but it's far from comprehensive. Today's web is more dynamic than ever and therefore more at-risk than it sometimes seems.

It is not just access to knowledge, but the knowledge itself that's at stake. Thousands of years ago, the Library of Alexandria was, as the astrophysicist Carl Sagan wrote, "the brain and heart of the ancient world." For seven centuries, it housed hundreds of thousands of scrolls; great works of philosophy, literature, technology, math, and medicine. It took as many centuries for most of its collections to be destroyed.

The promise of the web is that Alexandria's library might be resurrected for the modern world. But today's great library is being destroyed even as it is being built. Until you lose something big on the Internet, something truly valuable, this paradox can be difficult to understand.

* * *

Before the Internet, if you wanted to look up an old newspaper article, you usually had to find it in an archive. Which was how, one day in 1985, Kevin Vaughan found himself hunched over a microfilm reader, scanning *Denver Post* headlines from the winter of 1961. Vaughan, then a journalism student at Metropolitan State College, wanted to read an account of Skid Row on Christmas by one of his favorite professors.

"I was spinning my way through December," Vaughan said, "and I stopped and I saw this headline that said 20 children had been killed in bus-train collision. I can remember just staring at the screen and thinking, 'I've lived almost all my life in Colorado: How have I never heard of this?'"

After college, Vaughan became a reporter himself. When he covered a particularly gruesome train-crossing accident for the *Fort Collins Coloradoan*, his mind returned to the collision of 1961.

By this time it was 1992. One of the web's first browsers, Netscape Navigator, was still two years away from being introduced. *The New York Times* wouldn't launch its website for four more years. Google wouldn't be founded for another six. So Vaughan did what investigative reporters have done for years. He asked a police officer he knew if there was any way to find a man who didn't want to be found—the driver of the school bus from all those years ago. "He had some computer system right in the front seat of the squad car," Vaughan said. "He looked up the bus driver in the system and he jotted down his date of birth and his address and handed it to me. I carried around that scrap of paper for years."

Fifteen years, to be exact. Then, in 2006, as a reporter with *The Rocky Mountain News*, Vaughan's editors agreed to let him explore what had happened to the families affected by the 1961 tragedy in a series they would call "The Crossing."

The collision had new resonance for a community still aching from the 1999 Columbine High School massacre, a shooting in which 12 students and 1 teacher were murdered. "'The Crossing' grew out of Columbine," said John Temple, who was the editor and publisher of the *Rocky* at the time. "We realized that, as a newspaper, we would not be able to tell the community what the lifetime ramifications—on all the people who were in that school and were connected to that school—would be. Because it needed time to unfold. But this event gave us the ability to show how a single moment in time affects people for a lifetime."

Vaughan spent the better part of a year reporting the story. And in that time, a team of web designers, photographers, videographers, and engineers worked with him to

build a web experience around the series—the first time the *Rocky* had built something digital of this scope. "From a production perspective, it was a logistical monster," said Mike Noe, the interactive editor for the series.

It was worth the effort. Vaughan's story about the 1961 crash, a 34-part series that spanned more than a month in early 2007, was a sensation. "I don't want to overstate it," Vaughan said. "But I feel like it was transformative for the people who went through that tragedy. I've had people tell me, for instance, that the series cut loose all this emotion that they had bottled up inside, most of them for their entire lives." Readers wrote in to say they'd sit at their computers at midnight, refreshing their browsers until the next installment appeared. "It had tremendous impact," Temple said, recalling a community meeting that drew 800 people in response to the series. "It was a big deal."

"We did a couple of those public forums," Vaughan said. "In one of them, somebody asked John Temple how long the series was going to be on the Internet, and John said, 'Forever.'"

In 2008, Vaughan was named a finalist for the Pulitzer Prize in feature writing for the series. The next year, the *Rocky* folded. And in the months that followed, the website slowly broke apart. One day, without warning, "The Crossing" evaporated from the Internet.

* * *

What happened to the people of Greeley, Colorado, on December 14, 1961, was twice lost to time. To tell the story of "The Crossing" in the first place, Vaughan excavated a meaningful event—only to have the story he told require its own excavation. Before "The Crossing" was lost, before Vaughan could even tell it, he had to outline the filigree of events that traced back to a terrible and distant winter morning. The accident had been largely forgotten.

"It seems weird to think about something that killed 20 people having a very short period of time in the public consciousness," Vaughan said. There were newspaper accounts of the crash from 1961, but they raised as many questions as they

answered. There were allusions to public documents Vaughan would need to review—records that, for all he knew, could have been destroyed decades ago. In some cases, he got lucky. Like the time he was searching for a transcript from the trial of the bus driver, a 23-year-old man with a newborn daughter at the time of the crash. Vaughan wanted to read the man's statement, which was referenced in old newspaper clippings as having been long and dramatic, but never reprinted in full. A court clerk from the small county where the accident had happened agreed to accompany Vaughan outside town to an abandoned missile silo that had been converted into an archive of public records.

In 1994, there were fewer than 3,000 websites online. By 2014, there were more than 1 billion.

"I still think about it. It's almost dreamlike, surreal in my mind," Vaughan said. "You go in this door on the side of a hill and all of the sudden we arrive at this whole underground complex, which was built to withstand nuclear war." The room was the size of a basketball gym, and flooded with row after row of boxes, stacked up 20 feet or higher. It was, for the most part, a mess. There was no clear organizational system. So it was by chance that Vaughan caught a glimpse of a box, on a tall shelf far out of reach, with the last name of the prosecutor from the trial scratched onto the side. The clerk climbed a ladder to retrieve the box.

"First of all, the room looked like this scene in *Raiders of the Lost Ark*," Vaughan said. "I swear to God, I lifted the lid off this box and I felt like a shaft of light came down from heaven. I could see these big manila envelopes with the bus driver's name on them… Several dozen black-and-white photos taken over the course of the investigation that I had never seen before, a transcript of the court hearing held the day of the accident, all the letters the prosecutors got from across the country. It was just amazing."

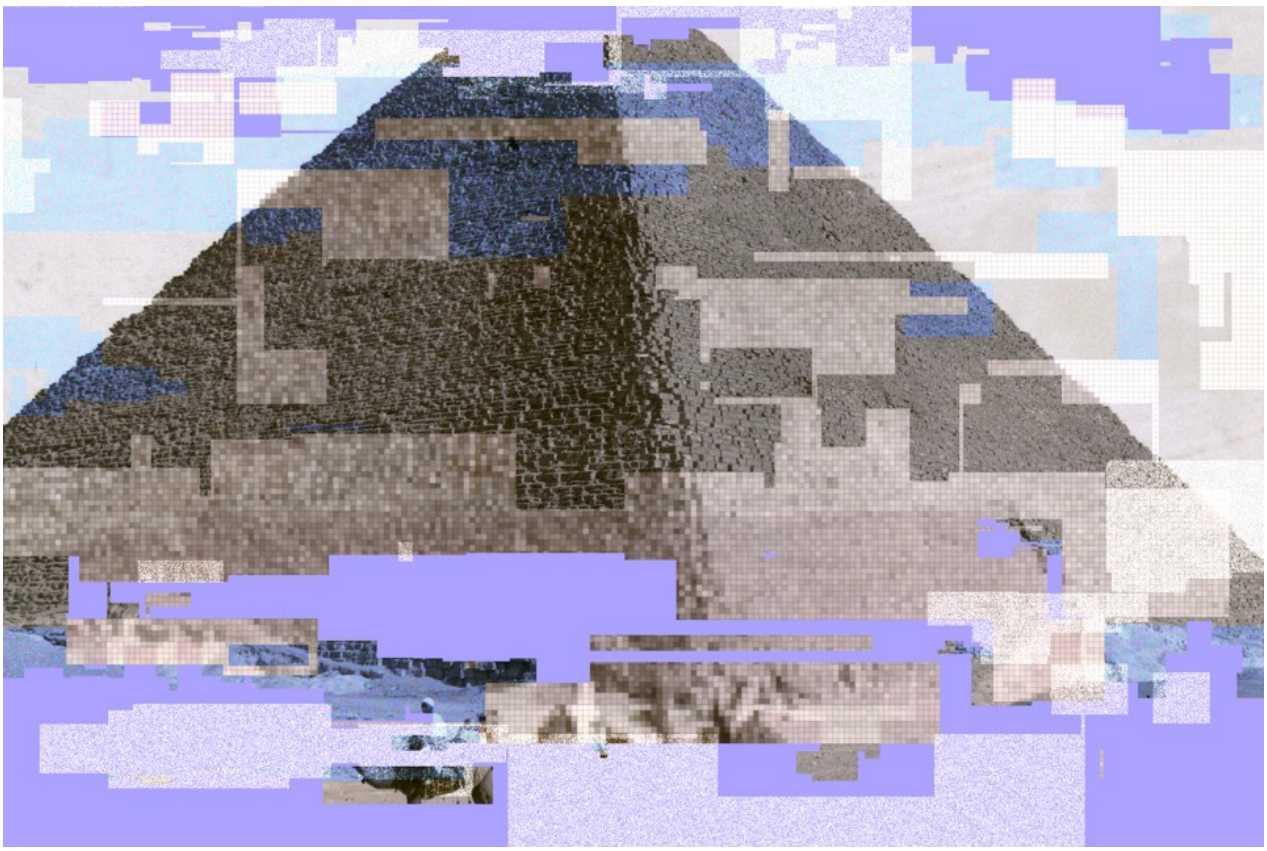
Later, after Vaughan had a chance to sift through the materials and make copies for the series, the clerk phoned the district attorney's office to ask what should be done

with the box. "They told her throw it away," Vaughan said. "She hung up and she looked at me and said, 'We are not throwing this away.'"

Many of the never-before-published documents and photographs Vaughan unearthed became key components of the web series, appearing only online and not in printed versions of the series. These weren't just extras, but key chapters of the story, told digitally. And when the website disintegrated after the *Rocky*'s closure, these stories weren't relegated to an old box on an unreachable shelf; they were gone.

"The day-in, day-out maintenance [of the site] just stopped happening, and so pretty quickly, some stuff didn't work," Vaughan said.

Vaughan began to think about how he might save the series. Was it even possible? "I wanted it up for a lot of reasons, but mostly I kept hearing in my head John saying, 'It'll be up forever.'"



Maison Bonfils / Library of Congress / Zak Bickel / *The Atlantic*

* * *

If a sprawling Pulitzer Prize-nominated feature in one of the nation's oldest newspapers can disappear from the web, anything can. "There are now no passive means of preserving digital information," said Abby Rumsey, a writer and digital historian. In other words if you want to save something online, you have to decide to save it. Ephemerality is built into the very architecture of the web, which was intended to be a messaging system, not a library.

Culturally, though, the functionality of the web has changed. The Internet is now considered a great oracle, a place where information lives and knowledge is stitched together. And yet there are no robust mechanisms for libraries and museums to acquire, and thus preserve, digital collections. The world's largest library, the Library of Congress, is in the midst of reinventing the way it catalogues resources in the first place—an attempt to bridge existing systems to a more dynamic data environment. But that process is only beginning.

In the print world, it took centuries to figure out what ought to be saved, how, and by whom. The destruction of much of Aristotle's work deprived humanity of a style of writing that the philosopher Cicero described as like "a flowing river of gold." What survived of Aristotle's writing wasn't prose, but more akin to lecture notes.

In other formats, entire eras of meaningful work have been destroyed. Most of the films made in the United States between 1912 and 1929 have been lost. "And it's not because we didn't know how to preserve them, it's that we didn't think they were valuable," Rumsey said. "The first 50 or 100 years of print after the printing press, most of what was produced was lost… People looked down on books as having less value in part because they were able to print things so rapidly and distribute them so much more rapidly that they seemed ephemeral."

Books, in their infancy, were trivialized the way the web is sometimes denigrated today. The telegraph was similarly maligned as "superficial, sudden, unsifted, too fast for the truth," as a critic in *The New York Times* put it in 1858. Transformative technologies in any era are met with initial skepticism, and that attitude often fuels indifference about initial preservation efforts. Historians and digital

preservationists agree on this fact: The early web, today's web, will be mostly lost to time.

Mostly, but not entirely. The Internet Archive has its Wayback Machine, an archive filled with imprints of web pages as they appeared in the past, like digital fossils. It's the closest thing we have to an online missile silo where folders can gather dust until the right person comes looking for them. "There's a school of thinking that says if you can't do it right, don't do it at all," said Scott, who began working for the Internet Archive in 2011. "But the thinking is, let's do what we can."

The life cycle of most web pages runs its course in a matter of months. In 1997, the average lifespan of a web page was 44 days; in 2003, it was 100 days. Links go bad even faster. A 2008 analysis of links in 2,700 digital resources—the majority of which had no print counterpart—found that about 8 percent of links stopped working after one year. By 2011, when three years had passed, 30 percent of links in the collection were dead.

More recently, a researcher at the Internet Archive has been running an analysis on the Wayback Machine to figure out what survives. "It won't be surprising to say that preliminary findings are showing things stick around for much shorter and changing constantly before they disappear," Scott told me. It is, as Jill Lepore wrote for the *New Yorker* earlier this year, "like trying to stand on quicksand."

The Internet Archive is getting more efficient. It can haul terabytes of data faster than ever—so fast, Scott says, that some servers end up shutting down early to avoid preservation. The 1990s web is mostly dead now anyway. Every once in a while, Scott will get word of a site founded around 1997 that's about to go under. "Those are really long-term tragedies," he said. "Simply because they're almost all gone."

Yet today's web is more at-risk than the iterations that preceded it. The serving environments are now more complex, and the volume of data involved is astonishing. In 1994, there were fewer than 3,000 websites online. By 2014, there were more than 1 billion.

"The interesting thing is that, at that time [in the 1990s], it was easier to archive the web because everything was flat web pages," said Alexander Rose, the executive director of the Long Now Foundation, an organization dedicated to establishing a framework for long-term thinking on a scale of 10,000 years. "So if you did save something, your chances of being able to see it and use it would be vastly better than if a company folded today, with deep back-ends of content-management systems like Drupal and Ruby and Django and all these things. The pages are not actual pages."

Saving something on the web, just as Kevin Vaughan learned from what happened to his work, means not just preserving websites but maintaining the environments in which they first appeared—the same environments that often fail, even when they're being actively maintained. Rose, looking ahead hundreds of generations from now, suspects "next to nothing" will survive in a useful way. "If we have continuity in our technological civilization, I suspect a lot of the bare data will remain findable and searchable," he said. "But I suspect almost nothing of the format in which it was delivered will be recognizable."

Fortunately, Vaughan had much of "The Crossing" saved to a DVD, which he'd used during guest appearances in college classes. "After the *Rocky* folded, I sent that disc to somebody at the *Denver Post* and asked whether everything was there that you would need to put this back online," Vaughan said. "The person looked at it and said, 'Yeah, actually, it is there.'"

* * *

Scholars believe that around 300 B.C., the Library of Alexandria may have housed three-quarters of humanity's texts. Today, three-quarters of humanity's books are abandoned, out of print and housed only in libraries—if at all. The existence of a resource, unfortunately, has little to do with access to it. That's true of today's libraries, too. The Denver Public Library acquired all of the print archives from *The Rocky Mountain News* when it folded. But only Vaughan had the source code that would be necessary to replicate "The Crossing."

So in 2009, the year the paper went under, Vaughan began asking for permission—from the library and from E.W. Scripps, the company that owned the *Rocky*—to resurrect the series. After four years of back and forth, in 2013, the institutions agreed to let Vaughan bring it back to the web. "I finally got the approval, and it was like, 'Now what?'" Vaughan said. "I don't know anything about building a website."

But Vaughan's son did. Two summers ago, when Sawyer Vaughan returned home after his freshman year at Olin College, they started talking about what it would take to make a replica of the series exactly as it had first appeared. "Sawyer said, 'Let me see this disc,'" the elder Vaughan recalled. "And he said, 'We can do this.'"

For Sawyer, most of the work involved combing through old code and adapting it for a today's web. In a pre-iPhone 2007, "The Crossing" had been designed as a desktop experience. It also relied heavily on Flash, once-ubiquitous software that is now all but dead. "My role was fixing all of the parts of the website that had broken due to changes in web standards and a change of host," said Sawyer, now a junior studying electrical engineering and computer science. "The coolest part of the website was the extra content associated with the stories… The problem with the website is that all of this content was accessible to the user via Flash."

Vaughan wanted the series to appear just as it did when it was first published, which meant still using Flash. It also meant attention to small but critical details. With Temple's help, Vaughan got permission from the designer Roger Black to use *Rocky*, the defunct newspaper's proprietary typeface.

Last month, eight years after it first appeared, "The Crossing" returned to the Internet. For Vaughan, and for anyone who reads the series, this is a success story. But it's also an anomaly.

"We've got the storage to be able to store all the books, all the movies, all the music, all the software, webpages, the whole damn thing," said Brewster Kahle, the founder of the Internet Archive, in a 2004 forum. "Everything, all the creative output of humankind, now can be in computers. And the Internet now can make it so that anybody that can access the Internet can access these things."

This is, after all, the great promise of the Internet: All of the knowledge, for all of humanity. A mix of legend and history tells us that the Library of Alexandria almost got there in its time. "We needed a technology change to be able to see this idea through again," Kahle said.

But the thing unsaid, the fact that unravels even an optimist's belief in what the web can be, is that the ancient library was eventually destroyed. Not by technology or a lack of it, but by people. Saving something and preventing its destruction are not entirely the same thing.

"At this point, if you mean the web when Tim Berners-Lee invented it, right now that web does not exist," Scott said. "Not really. News organizations kill old articles, YouTube's old videos go away. And while the Archive and other entities are saving—quote-unquote *saving*—these sites, even those will go to new URLs. They won't be in the same place. You'll have to search for them... There are success stories. But meanwhile, silently, thousands of useful things are disappearing. As time goes on, I have even less and less hope for how long it will last."

ABOUT THE AUTHOR

ADRIENNE LAFRANCE is the editor of TheAtlantic.com. She was previously a senior editor and staff writer at *The Atlantic*.

Twitter Facebook