The Washington Post

Old Dominion U. professor is trying to save Internet history

By Daniel de Vise, Published: July 17

What if you woke up tomorrow and all of your painstakingly edited YouTube videos were gone, your 4,000-entry Twitter feed erased and your lovingly tended Facebook page deleted?

Michael Nelson, a computer science professor at Old Dominion University in Virginia, is thinking those terrible thoughts. His research team has spent the past couple of years studying how much of the Internet is being saved — what portion of the vast sea of online ephemera is preserved in some permanent archive.

Nelson is an Internet time traveler, one in a small community of scholars and techies who are laboring to build a past for a technology obsessed with the present. In a computing culture accustomed to deleting its Internet history, they are trying to create one.

“We’re sort of stuck in this perpetual now,” Nelson said. “Figuring out what was on the Web an hour ago, a day ago, a week ago, we’re really bad at that.”

Nelson and some colleagues at Old Dominion and the Los Alamos National Laboratory have developed a sort of Internet time machine called Memento . When attached to a browser, it enables the user to search for a Web site as it appeared on some past date, if an archived page exists.

Joseph JaJa, a professor at the Institute for Advanced Computer Studies at the University of Maryland and a fellow time traveler, is working on another tool that would allow a search of the archived Internet as it existed at a time of one’s choosing.

“The Internet now is the main communication and publication medium,” JaJa said. “If we don’t preserve it, we lose a good part of our cultural heritage.”

Computer users who think their Flickr photos and Facebook updates last forever could be in for a shock. The average life of an Internet page is about 100 days. When Nelson’s team surveyed users about lost Web sites, they found many causes: service providers bought and sold; servers seized by police; page owners dying, leaving for college or simply losing interest.

Remember GeoCities? The community of user-designed pages — some termed it the Facebook of the 1990s — was shut down in 2009. Yahoo Video, a onetime YouTube rival, closed to user-generated content last year.

All of this runs counter to the notion that anything posted online, particularly if it is unflattering, is permanent. That is not true — although highly publicized online gaffes tend to endure because they are so easily copied.

Much of what has been published in the roughly two-decade history of the Internet is eminently disposable: 140-character musings on the weather, colorless corporate directories, personal ads and a seemingly endless photographic celebration of cats.

Yet scholars are growing concerned about the burgeoning quantity of creative work — Twitter aphorisms and blog posts, photographs and videos, even scholarly papers — that is "born digital," without corporeal form and doomed to die online if it is not salvaged.

Future historians might want to study today's online flat-stomach ads in the same way contemporary scholars ponder cigarette ads from magazines of the 1960s as a barometer of culture. Internet coverage of the Sept. 11, 2001, terrorist attacks may prove as historically resonant as TV coverage of President John F. Kennedy's assassination.

And biographers will be hard-pressed to chronicle President Obama's profoundly digitized 2008 campaign without archival images of www.barackobama.com.

But much of that heritage is lost.

The Web, the global network of documents connected by the Internet, went online sometime after 1990. That is not so long ago. Yet, nearly all of its early content is gone because no one thought to preserve it.

"It was conceived without the notion of time and without the notion of archiving at its core," said Herbert Van de Sompel, a computer scientist who works at Los Alamos and collaborates with Nelson.

The Internet Dark Ages ended in 1996, when Brewster Kahle, an entrepreneur, began preserving Web pages by the billions in the Internet Archive.

Every two months, Kahle's nonprofit library dispatches a computer program that crawls through the Web and stores every page it finds, except those whose owners don't wish to be found. Today, the archives hold 3 petabytes of information, which is the numeral three followed by 15 zeroes (there are a million gigabytes in one petabyte) — and it is one in a network of archives around the globe.

"Whoever is going to be president in 2048, she's in high school now, and she may have a Web site, and we probably have it," Kahle said.

The Internet of today is effectively infinite: a universe of more than 1 trillion unique pages, expanding by 200 million tweets every day and by 24 hours of YouTube video every minute.

One cannot divide by infinity. So, to estimate how much of the Web was being saved, Nelson and his colleagues took a sampling of 4,000 Web pages from four sources.

Their findings are messy but instructive. When Nelson's team tracked Web pages chosen with search engines and selected more or less at random, it found that only 19 percent had been archived. When it tracked pages from Delicious, a social bookmarking site akin to Digg, it found that 68 percent had been preserved. Pages harvested from bitly, an address-shortening site, were less likely to be archived. But most pages taken from the Open Directory Project, a public index of Web sites, were saved for posterity.

The lesson: Popularity on the Web equals longevity. If your Web page has been bookmarked or indexed, it has received a measure of recognition and is more likely to endure. Preserving what is important now is easy: Important things are copied and shared, again and again. But what about things that will become important later? Twitter feeds from a yet-to-be-famous author or the YouTube offerings of the next Spielberg?

"Let's assume that 99 percent of what's on the Internet is" junk, said Matthew Kirschenbaum, an expert on technology in the humanities at U-Md. "That still leaves 1 percent. And if you think of how big the Internet is, even that 1 percent is a very big deal."