Why the United States Needs Digital Preservation Policies

Matthew Scott Weber, Rutgers University-New Brunswick

Digital information is fragile, especially when compared to printed products. Content on the Web is easily modified, edited, and deleted. And as recent developments have underscored, content on the Web is also easily faked and falsified. The average lifetime of a website is less than three years, and a recent study found that within two years, nearly half of all links on web-based content point to pages that no longer exist. This means that content published online is lost at an alarming rate, and is generally not preserved anywhere.

In the digital era of fake news and misinformation, one critical and often overlooked issue is the need to maintain a common record of digital content. This issue — commonly called data preservation or digital preservation — is inherently complex. It is complex because the volume of digital data produced today is massive and difficult to comprehend and because it is challenging to clearly spell out what should be preserved. Nevertheless, with an increasing majority of information being produced digitally, a national data preservation policy is needed to preserve both the history of the U.S. government and the integrity of the nation's collective national data.

What Should be Preserved?

Unlike other countries, the United States does not have a traditional data preservation policy. One reason for this is that it is nearly impossible to create within U.S. boundaries an archive of the World Wide Web and the related born-digital content that is produced and managed digitally. The United States does not have a dedicated Web domain like many other countries do, including France (.fr), Israel (.il), Egypt (.eg), China (.cn) and others. Most commercial content in the United States is published on the .com domain, which does not have a geographic constraint. Likewise, U.S. educational institutions use the .edu domain, which is not restricted to a single country.

Some domains, however, are easier to define. For instance, the .gov domain is designed for government websites. It would, therefore, be possible to create an initial sample of digital content for preservation that would establish a starting point for future digital preservation efforts. To that end, a number of preservation efforts are already underway within the United States public sector.

Current Efforts in the Federal Government and Beyond

The web was born in 1989, but it wasn't until 1996 that the Internet Archive began to preserve small batches of Web based content. The Internet Archive has gone on to be the largest repository of born-digital content in the world.

The Library of Congress has been archiving born-digital content since 2000. The Library focuses on the preservation of content pertaining to the U.S. government, as well as information by and about candidates running for office, political commentary, information about political parties, media, religious organizations, and other groups. In addition, the Library of Congress has launched a number of broader initiatives to fund web archiving of topical areas such as "business and technology," but these efforts are sparse and focus on sampling data.

From 2000 to 2016, the Library of Congress' National Digital Information Infrastructure and Preservation Program sought to advance best practices for digital preservation. Today, the Federal Agencies Digital Guidelines Initiative is a collaborative cross-agency effort to develop guidelines and best practices for born-digital content. The guidelines do not provide specifications about how to preserve born-digital content, but represent an attempt to articulate how content should be produced for digital environments.

In addition to the federal efforts, the End of Term Web Archive is a project run by the California Digital Library and the Internet Archive. This archive aims to capture and save U.S. Government websites as they existed at the end of each presidential administration. The project began in 2008, and has continued to the present day. The project has a multitude of partners, including the Library of Congress.

Comparatively, the United States is far behind other countries in terms of digital preservation efforts. Denmark enacted the Legal Deposit of Published Materials Act in 2005, requiring the collection and preservation of Denmark's .dk Web domain. The British Government maintains a robust archive of U.K. government websites, as does Australia, France, Taiwan, Japan, and Portugal, among others.

Ethical Implications of Preservation

In addition to the challenges of scale, pace of change, and the other technological challenges, there are extensive copyright challenges associated with web archiving. Outside of the government domain, the majority of content on public websites is owned by third parties. The New York Times, for example, owns the copyright to news articles published on its websites. Without a mandate to archive or changes to the U.S. Copyright Act, it is unlikely that Times web content could be preserved without permission.

Social media content is especially challenging. Companies such as Facebook and LinkedIn, have unique digital ecosystems within which content is produced, stored, and accessed. So there are added barriers that prohibit outside entities from archiving and preserving content within those domains. Yet such content is critical for understanding today's society and politics.

Next Steps

Ultimately, works of the U.S. Government are not necessarily subject to copyright protection, so digital preservation can start there. To support this, a clear policy is needed to insure that digital content produced by the U.S. Government is accurately preserved and made available as a public record. Creating such an archival record could help combat misinformation and fake news — and serve as a model for broader digital archiving reaching beyond government in the future.

Read more in Matthew S. Weber, "The Tumultuous History of News on the Web," in Niels Brugger, & Ralph Schroeder (Eds.), *The Web as History: Using Web Archives to Understand the Past and the Present,* (UCL Press, Forthcoming).