University World News

UNITED STATES

Project launched to improve web archiving worldwide

13 July 2018 Issue No:514

Virginia Tech is leading a project to make web archives more valuable to researchers worldwide by developing course materials and cyber infrastructure to teach librarians and archivists internationally how to collect, extract and analyse archived information from the World Wide Web.

The Institute of Museum and Library Services recently awarded a US$248,450 grant for the collaborative two-year project, Continuing Education to Advance Web Archiving, which ultimately aims to help archivists use new tools and capabilities to make more effective use of web archives.

Zhiwu Xie, director of digital library development for the University Libraries at Virginia Tech, is leading the team of library and archive experts to create a curriculum surrounding the technology of web archiving and challenges related to how archivists and librarians can gather the most useful information from archived internet sites and social media.

“The web is the most prominent channel of communication we have today, and web sites change all the time. The web doesn’t have a memory, so a history of time is hard to construct,” said Xie. “Web archiving is about recording that memory.”

Project team member and Virginia Tech Professor of Computer Science Ed Fox believes in providing individuals and libraries with the tools to better access and analyse the massive amount of information already archived.

“I view information as a fundamental need of humans,” said Fox, who also serves as the director for the Digital Library Research Laboratory. “The most visible information is what’s available over the World Wide Web, and over time, in its archive. This information is invaluable for researchers studying areas such as trends in elections, technology and the environment.”

More than 10s of petabytes of web content have been collected and archived by memory institutions. All of the project collaborators are pioneers in web archiving technology and infrastructure. They include Xie, Fox, Martin Klein from Los Alamos National Laboratory, Michael Nelson from Old Dominion University, and Justin Littman

from George Washington University, all in the United States; Ian Milligan from the University of Waterloo, Canada; and Jefferson Bailey from the US-based non-profit archiving organisation Internet Archive.

Making an impact

“Collectively, we have done a lot of work in creating tools for web archiving; we want to put our work to use and make an impact on society,” said Xie.

“By creating training materials for some of the most innovative and complex tools used in web archiving, it can help lower barriers for institutions wanting to run these technologies locally, either for collecting, or especially, for researcher and user support,” said Bailey, who serves as director of web archiving at Internet Archive.

“Suites of open source tools are available to assist researchers conducting analyses and extracting knowledge,” said Xie.

“However, these tools require the user to be proficient in big-data processing and analysis. Very few librarians or archivists have been trained to understand, use, maintain and manage these tools.”

By the end of the project, the collaborators will provide a collection of educational resources, a series of in-person and online training workshops, and cyberinfrastructure for deploying tools to support the curriculum and workshops – including source code.

“The curriculum will include project-based learning because people learn better by doing,” said Fox. “During the training, participants will solve problems like they would face while helping patrons. The curriculum will be need-oriented as opposed to system or technology oriented. All of the training and tools will be free to the user.”

“By educating more people and organisations on the technologies of web archiving, the project can contribute to allowing more organisations to build collections of web-published materials,” said Bailey.

“This benefits society by ensuring a greater portion of web-published historical documentation is preserved and accessible into the future.”

“Equipped with these skills, library and archive professionals will be able to go beyond their traditional role as information providers or pointers and form deeper alliances with researchers,” said Xie.

“This will continue to transform libraries and archives from information repositories to knowledge producers.”

Virginia Tech mass shooting a catalyst

Located in Blacksburg, Virginia, Virginia Tech was the site of one of the deadliest mass

shootings on a college campus, when in 2007 an undergraduate student armed with semi-automatic pistols shot 49 people, killing 32 and wounding 17.

Fox, who joined Virginia Tech in 1983, said the shooting was a catalyst for his work on web archiving. “As computer scientists working with collection and managing information, I asked the question, when something like this happens, what can I do to make the world better?” he told *University World News*.

“So we shifted efforts to create and aid use of archives of information that would otherwise be lost. It began with the shooting here and went to crises and tragedies elsewhere and now we do all of that and analyse trends to help understand emergencies and global changes.”

He says providing a curriculum to train more people is vital because web pages change so much faster than traditional forms of information.

“We are constantly losing history, so it is important for people to be aware of these issues and more skilled and knowledgeable about building and using archives. We need to get more people involved in the process.”

With the emergence of digital humanities, wherein researchers are more involved with using computer methods, more and more people in the academic world are able to use the web tools for archiving.

But there are thorny problems to deal with due to the changing nature of information on the web. One of those is how to deal with information that has been put out there but since retracted.

“The consensus is that you have to keep both the original version and the corrected version in web archiving – unless there is legal action to remove them. But the broader area of retraction includes how to get the word out about each incident,” Fox says.

“There are people doing research into this but if we don’t archive comprehensively the whole thing becomes moot.”

Losing history is another challenge, for instance if publications close down, taking their archive with them into oblivion. There are some existing projects such as LOCKSS, based at Stanford University Libraries, California, which work with publishers to keep copies of journals librarians have subscriptions to and protect them even if they go out of business.

They can involve creating a dark archive where copies are kept but access is limited to those with certain rights.

“There are research tools, although not for public use. Part of what we do is put them in a cloud in a way that people can use them or through courses get experience of using

them. And one thing we are doing is making them more robust so that other people can use them more easily," Fox said.

http://www.universityworldnews.com/article.php?story=20180713162517805