Data may not compute: Program stores older Web research files, left at risk by technological leaps

September 19 2011, By Alvin Powell

"Data is not like a book. If you get a 300-year-old book and you know the language, you can usually read it," said Gary King, the Albert J. Weatherhead III University Professor and head of Harvard's Institute for Quantitative Social Science (IQSS). "Data changes formats. If it's from even five years ago, you might not be able to read it at all." Credit: Kris Snibbe/Harvard Staff Photographer

Modern scholars are wrestling with a problem that ancient monks and early authors managed to master: how to keep their work accessible to future generations.

While the books, papers, and journals of early scientists remain readable to anyone who can lay hands on them and knows the language, and that is not the case for those whose work is stored on early computer media,

just a few decades old.

The breakneck pace of technology's advance has left data in its dust, stored on tapes, floppy disks, and other media now unreadable by newer computers. And it's not just the nature of storage media that is rapidly changing. File formats change as new programs are developed, rendering older programs obsolete even while giving researchers powerful new tools.

"Data is not like a book. If you get a 300-year-old book and you know the language, you can usually read it," said Gary King, the Albert J. Weatherhead III University Professor and head of Harvard's Institute for Quantitative Social Science (IQSS). "Data changes formats. If it's from even five years ago, you might not be able to read it at all."

King has watched those changes since he arrived at Harvard in 1987. As head of the Harvard Data Center, then the Harvard-MIT Data Center, and now the institute, King realized long ago that efforts had to be made to ensure access to digital data for future scholars.

While publication in academic and scientific journals provides summaries of research, King said those articles are like advertisements for the underlying work, the reams of data gathered during exhaustive social science surveys, years of field observations, and long nights in the lab. Further, he said, today more grant-making agencies and journals require researchers to make their data available to others as a condition of a grant or of publication.

"It's very important in science and social science to share research data," King said.

One solution to the problem already exists, on computers at Harvard and in a growing network of corporations, universities, and other institutions.

Called the Dataverse Network Project and spearheaded by the IQSS, the effort provides archival storage for research projects, initially in the social sciences but recently expanding to the physical sciences and humanities.

The Dataverse project solves problems that plagued the two most common previous data storage strategies, King said. The first is that researchers sometimes use major archives to hold their data. The problem with that, King said, involves loss of control over the data and, potentially, a loss of credit for gathering it, because the archive is sometimes cited as the source. The second commonly used strategy is to store the data on personal computers or servers, making it available on the Web through a researcher's Web page. The problem there, King said, is that Web pages don't endure for long. Researchers change institutions, links are lost, and access to data is gone as well.
"The average age of a link on the Web is very short," King said. "Servers under the desk break or are replaced; the data can disappear."

The Dataverse project is designed to solve both problems, King said. First, the IQSS employs professional archiving standards that ensure access to data long into the future. Once a researcher's data is put into the system, it is converted from its original file format into a basic one that ensures the information will remain readable for decades to come. When that format becomes obsolete, King said, the system will automatically convert it to a new format, also designed to endure for decades. To guard against loss, the data is backed up on servers at different locations.

Instead of being locked away somewhere, the data remains accessible to the researcher through a Web interface designed to look like just another page — holding a list of datasets — on the researcher's website. Instead of bringing visitors who click on a page to a researcher's server, though, it links directly to a Dataverse server. The data sets, like the journal

articles that result from them, have their own citations so that, if they are used by other scientists, a researcher gets credit for the work.

“As a researcher, I don’t need to do anything. It looks like it’s mine, but it’s preserved in the background,” King said.

There are Dataverses at several different levels, including the Dataverse Network Project, which has developed and distributed the software; the IQSS’s Dataverse Network, which is the Harvard-centered network, holding the data of Harvard researchers; the Dataverse networks of other institutions; and the Dataverses of individual researchers, which are individual archives from their specific projects and which reside on the networks at specific institutions.

Mercè Crosas , director of product development at IQSS, led the development efforts of the Dataverse Network software. She said IQSS currently hosts more than 350 individual researchers’ Dataverses. Those Dataverses hold about 40,000 studies, made up of 665,000 files. Although Dataverse has so far mainly been used by social scientists, Crosas said some groups in the sciences, including the Harvard-Smithsonian Center for Astrophysics, are beginning to explore Dataverse options.

She expects the size of the files stored there to double in the next five years, as more researchers seek solutions to the problem of storing data into perpetuity. To help that expansion, she said, the Dataverse software is open source, meaning that the code is open to others to download and edit. Among the institutions that have adopted the Dataverse approach are the University of North Carolina, the University of Michigan, and several campuses of the University of California.

The software’s open-source nature means that other institutions can have their own programmers add features that can then be shared with the

community of users.

Of course, preserving anything into perpetuity is a tall order, and King acknowledged that will be a central challenge as people and institutions change. The advantage of a place like Harvard, though, is that it is stable and likely to endure.

“You need the community to persist,” King said. “That’s the kind of thing Harvard does best.”

This story is published courtesy of the Harvard Gazette, Harvard University’s official newspaper. For additional university news, visit Harvard.edu.

More information: dvn.iq.harvard.edu/dvn/

Provided by Harvard University

Citation: Data may not compute: Program stores older Web research files, left at risk by technological leaps (2011, September 19) retrieved 2 May 2024 from https://phys.org/news/2011-09-older-web-left-technological.html