http://www.bostonglobe.com/ideas/2015/08/10/computer-scientists-have-looked-for-solution-that-doesn-exist/tXO0qNRnbKrClfUPmavifK/story.html?p1=Article_Trending_Most_Viewed

The Boston Globe

The race to preserve disappearing data

By Bina Venkataraman MAY 17, 2015

CASABLANCA" AND "Citizen Kane"are exceptional not just for their wide appeal: American feature films shot before 1950 faced about 50/50 odds of surviving into this century. Countless independent and documentary films from the first half of the 20th century, and about 80 percent of the mostly silent movies made in the 1910s and 1920s were lost, according to the Library of Congress, due to either neglect or studios that set them aflame.

All that changed with more stable — and less explosive — film stock and a sense among Hollywood's major studios that preserving films was in their interest. Indeed, those films could be resold as home videos, then as DVDs, and now per streaming view; some even grew cult followings years after release. Tucked away in temperature-controlled vaults, master reels of

classics like “Dr. Strangelove” and “Star Wars” have a high chance of surviving for hundreds of years.

But now that filmmaking has gone digital, a new threat to cinema history has surfaced. It’s born of a paradox: The same technology that has drastically driven down the cost of filming and editing creates obstacles for preservation. And while these technologies have unleashed the voices of independent filmmakers who might have never made movies in the analog era, it has put at risk future generations’ ability to watch them.

Vint Cerf, a founding father of the Internet, warned at a recent meeting of the American Association for the Advancement of Science that our era is in danger of becoming a “digital dark age,” from which we’ll leave our progeny too little information to grasp their history. But who is responsible for bearing the cost of preserving digital artifacts and knowledge for the future? And how do we determine what’s worth saving?

THE IRONY of Cerf’s concern is that the digital age is anything but dark. We are in the era of big data, exploding with exponentially more bits and bytes each year. By one back-of the-envelope estimate, the number of digital photos we snap in two minutes exceeds all the photographs taken during the entire 19th century. Faster computing speeds; sensors on our phones, cars, and transit systems; and falling costs of technologies to sequence genomes and launch satellites contribute to the data deluge. We’re entering the era of the “Internet of Things,” in which virtually any object or organism on the planet could one day collect and transmit data.

Meanwhile, both the cost and ease of digital storage — at least for the average person — have plummeted. Cloud storage offerings from companies like Amazon, Google, and Apple have freed photos and documents from the perils of floods or fires, but digital files can still vanish into the ether. The cloud isn’t yet robust enough for long-term archival of complex datasets and gigantic master movie files. Nor can it keep up with predicted demand. The International Data Corporation projects that in five years, storage capacity will match just 15 percent of the data in the “digital universe.”

The challenges of digital preservation worry film archivists and aficionados alike. Three out of four features screened in American cinemas are independent films, according to a 2012 report from the Academy of Motion Picture Arts and Sciences. Yet most independent filmmakers have neither a plan nor budget to preserve their films. A rare few are bought by studios for wider distribution and preservation or adopted by nonprofit film archives. More get licensed temporarily, leaving no one on the hook for long-term preservation. At risk is the archival footage of the future. Imagine Ken Burns documentaries without footage of 20th-century baseball games or of Prohibition-era speak-easys.

Digital preservation is essentially a hot potato problem, where everyone wants to pass responsibility onward.
Dr. Francine Berman, chair of the Research Data Allianc

“There is no way of ensuring the lifespan of a digital image sequence,” says Dave Diliberto, a veteran film editor and producer who has worked with the Coen brothers and Errol Morris. “Everything you can do to back it up is expensive, and no one knows how long the files will last.” Major Hollywood studios, Diliberto points out, are now spending large sums to shore up born-digital movies on old-school film, but

The problem of preservation is not unique to the film industry. It spans the digital artifacts of our age — from photos to music to scientific research data. One study of more than 500 biology papers published from a 20-year span found that as time passes, less original research data can be found; it suggested that up to 80 percent of raw data collected for studies in the early 1990s is lost. A crucial virtue of science is that researchers can reproduce findings or correct them over time by reevaluating original data. Fields from epidemiology to education to climate change require records that span decades or longer.

Lost data also plagues the legal world. A 2013 study of Supreme Court decisions by Harvard University Law School professors found that so-called link rot is eroding intellectual foundations of legal scholarship: Nearly half of all Supreme Court decisions up to that date and

more than 70 percent of law journals from 1999 to 2012 referred to Web pages that no longer existed.

NOT EVERYTHING is worth saving for the long haul. There are billions of e-mails sent in a week for transient work, daily digital to-do lists, and measurements collected every second by sensors on planes and power grids. Much of the data we create are useful for a moment but quickly age into digital detritus with unclear value for future generations — not unlike old shopping lists or blurry photographs stored in shoe boxes — only now, in far vaster amounts. The digital era exalts the instantaneous, but the archivist's perspective is long. We snap photos and send e-mails thinking of now, not how distant generations might view or read them. Captains' ship logs from the 19th century, for example, have proven instrumental in tracking historical weather patterns that now help us understand the changing climate. A home remedy found in a ninth-century Anglo-Saxon manuscript has recently shown potential to fight drug-resistant staph infections.

Independent filmmakers surveyed by the Academy had two main concerns: making sure audiences saw their current films and moving on to the next movie. More than 60 percent said they do not migrate their digital files to new formats, a step to create access over time — even though they and their heirs could retain rights and earn royalties on the footage for nearly a century. "Independent and documentary filmmakers need a budget for preservation," says Andy Maltz, managing director of the Academy's Science and Technology Council.

What was once a race to rescue information from going-extinct media (think of old files trapped on floppy disks) has morphed into a mounting need to copy and curate massive troves of data, says Dr. David Rosenthal, the founder of a library-led digital preservation network run out of the Stanford University libraries. Digital information decays over time and files grow corrupt from "bit rot," which Rosenthal says is best fended off by creating copies of data in multiple virtual and physical locations.

Even if we could save all data in the cloud, we would have to invest in annotating data to create searchable archives that save useful knowledge for people in the future. Preservation requires paying for back-up storage sites and media, and actively managing data at intervals, as opposed to just storing and ignoring files or film reels, says Maltz. By what Maltz acknowledges is a slightly dated estimate, the costs of digital archiving of a film can be up to 11 times more than the cost of archiving an analog film. Digital films, like symphonies, have many constituent parts — audio, video, subtitles — which all need to work to make them play.

Saving scientific data poses similar economic challenges. In 2013, the president's science adviser, Dr. John Holdren, directed federal agencies to call for the scientists they fund to make their raw data publicly available. Several journals have also begun requiring that data be published along with findings. The challenge is how to meet these requirements as datasets balloon in size — not just at the moment of publication, but over time. In a commentary in Science, Cerf and Dr. Francine Berman, chair of the Research Data Alliance, wrote that this call to make research data accessible has not been matched by public or private sector funding for the infrastructure needed to host and preserve the data. Without ongoing investment to curate research data, tag its relevant context, create copies, and maintain access, stored scientific data could become lost or meaningless.

"Digital preservation is essentially a hot potato problem, where everyone wants to pass responsibility onward," said Berman, also a professor of computer science at Rensselaer Polytechnic Institute. She notes that in the private sector, companies invest in preserving data that give them a competitive advantage. The larger challenge is preserving those digital artifacts that have broad societal relevance for the future, but no urgent private interest.

Publicly funded archives such as the National Archives and those supported by federal R&D agencies fulfill only a fraction of the preservation needed to pass on society's knowledge to the future. Less than 1 percent of the Library of Congress's 1.4 million archived videos and film reels were born digital. While the Library of Congress can preserve digital films if filmmakers share

their unencrypted files, less than a dozen filmmakers and studios have done so, and the library has yet to preserve a single born-digital feature-length film.

Philanthropic and nonprofit efforts have emerged to help bridge the gap. The Long Now Foundation, for instance, is working to preserve all of the world's languages on a nickel plate dubbed the Rosetta Disk. The Internet Archive has a machine that crawls and copies the Web regularly — albeit imperfectly, without backing up content that lies behind paywalls and other barricades. Public-private partnerships to archive data have also emerged, such as the Alzheimer's Disease Neuroimaging Initiative, supported by the National Institutes of Health and several pharmaceutical companies. Scholars have launched nascent efforts to address the link rot from the references sections of law journals and court cases.

While digital technology can distract us from the long term, it can also offer more automated ways to save our relics, says Jeremy Leighton John, an investigator for the Digital Lives Research Project at the British Libraries, noting that diaries once left to a single heir can now be copied and propagated, what he calls an emancipation of personal archives. And it is sometimes possible, with enough money and interest, to raise lost digital creations from the dead. Last year, students at Carnegie Mellon University and the artist Cory Arcangel resuscitated from floppy disks works that Andy Warhol created on an Amiga computer in 1985.

Bina Venkataraman is director of global policy initiatives at the Broad Institute, a lecturer at MIT, and former senior adviser for climate change innovation in the Obama administration. Follow her on Twitter @binajv.