Virtual Archaeology Review

Digital Conservation and Access: Saving Humanity's History in the Petabyte Age

Michael Ashley

Cultural Heritage Imaging, USA

Abstract

We are at a unique point in history, the cusp of a Digital Dark Age, where cultural heritage professionals must work to care for the physical past while assuring that there will be a digital Rosetta Stone for future generations. This contribution describes the state-of-the-field in digital preservation and access, and is a call to action for individuals and institutions alike to work beyond our comfort zones and competitive boundaries in order to help define a sustainable digital future. Defined as an "hourglass of participation", I describe a method where knowledge producers, curators and consumers interact and actively work to make content born-archival and long-term viable, semantically managed and ready for reuse and public dissemination.

Key words: *DIGITAL DARK AGE, PETABYTE*

1. The Peril

This is about the sharing and preservation of human traces digitally, and coherent access to these traces. It is really about time, memory, and perception, the persistence of history. This is a call to action for individuals and institutions alike to work beyond our comfort zones and competitive boundaries in order to help define a sustainable future as we enter into the 'Petabyte Age.' Here, we will look at the state-of-the-field in digital archiving and access to see if the world is ready for such innovations, and where it isn't, what we can do about it.

> "Thousands of years ago we recorded important matters on clay and stone that lasted thousands of years. Hundreds of years ago we used parchment that lasted hundreds of years. Today, we have masses of data in formats that we know will not last as long as our life times. Digital storage is easy; digital preservation is not."
> - Danny Hillis (Brand 2003)

How is it possible that the human commerce of the now, our arts and sciences and creativity and histories, are entrusted, bound and embedded in media that are ephemeral and fragile, written in formats our descendants won't be able to decipher, if they can read them at all? Will there be a digital Rosetta Stone? In actuality, as Hillis points out, this is our state of the field, and what we should be asking before things go more wrong, is, 'What are we going to do about it?'

The Internet is not ether, it is housed in cables, powered by electricity, built, maintained and co-created by millions of people worldwide. Some 1.2 billion people are depending on it. In February 2008, an undersea cable that was severed through age and neglect cut off Internet capabilities for much of the Middle East, North Africa and Asia (CNN 2007). Imagine no financial transactions, full blackout of countries, and cultures out of global communication. This week, China and Russia were accused of cyber-espionage for planting devices that can take over the Unites States power grid (Gorman 2009). As I write this, the Internet has been cut by vandals, blacking out whole parts of Silicon Valley (Gomez 2009).

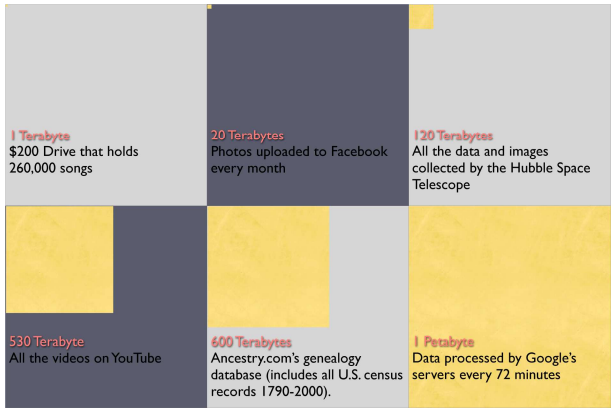

Figure 1: The Petabyte Age

The Internet, and digital technology, remains volatile, friable and at high risk from the perspective of long-term human history. Digital technology is radically changing how we produce knowledge and interact with each other, and not necessarily always for the better. The tech industry measures time in financial quarters and product lifecycles. Those of us who care about the future of human knowledge need to step up and figure out how to make digital content persistent, insulated from the sea changes of innovation and stock prices. This is, as Stewart Brand says, a "civilizational issue" (Brand 2003).

2. The Promise and the Price

What if we could store the shared corpus of human history digitally, make it safe, durable, and secure? What would be the impact of such a capability for sharing, understanding, and research? We are seeing the potential of this ideal being realized today. Never before in human history has it been more possible to share our knowledge globally, instantly, and with indescribable impact.

In an article in the 16.07 issue of Wired magazine, the author argues that the scientific method has been made obsolete because the human condition is now crawlable by grid computers driving phenomenally powerful algorithms (Anderson 2008). Servers, such as those run by Google, can process a petabyte of information in just over an hour. That's 1,000,000,000,000,000 bytes, a million gigabytes or 4 million 4-drawer filing cabinets of documents. Such power comes at a price, and one such price is the tangibility of data. Human knowledge is no longer stored routinely on media we can archive, but in 'cloud' computers, distributed systems of interdependent servers, networks, and human operators. I will return to the quite tangible challenges introduced when corporations are the stewards of human meaning.

This foundational layer, algorithmic processing at the petabyte level, is the basis of Google's efforts, and while it is compelling, it is also highly controversial. And rightfully so, as some content should not be shared, such as medical records or sensitive archaeological information, and other content is not ready to be shared, such as unpublished or raw work. This said, making our digital universe safe, secure and accessible is something we can probably all agree is a nice Utopian vision.

Clearly (at least I think it is clear), institutions of cultural memory and individual contributors and researchers continue to have vital roles in the gathering, creating and sharing of digital content. We are both the producers and consumers of content. We are the producers and consumers of the Algorithm.

3. Architecting Participation and the Hourglass of Sharing

We want to believe in Google's mission to "to organize the world's information and make it universally accessible and useful." We believe that to bring this mission to fruition requires direct intervention from people at every step of the workflow - from initial idea brainstorming through to archiving, publishing and remixing, and indexing. The better the data, media (we'll call it content) are, the better the algorithm, the more meaningful the human cloud computer will be. Our findings show that the barriers to sharing are generally cultural rather than technological. In the programs described below, we are working to make it as easy as possible to share from a technical and institutional standpoint by rewarding the act of sharing with subsidized digital preservation, and by demonstrating the value of contextualized, shareable content.

Open Knowledge and the Public Interest (OKAPI) brings together faculty, students and staff at the University of California, Berkeley, to promote open knowledge and free culture on campus and around the world. OKAPI's primary goal is to forge new tools for open learning and collaboration across borders and communities (Wittman 2008). OKAPI partners with national and international educational, scientific and cultural organizations to share knowledge and expertise.

We have devised an "hourglass of participation", in which we have identified a sweet-spot where knowledge producers, curators and consumers interact and actively work to make content more digitally durable, reduce intellectual property constraints, and prepare this content for reuse and public dissemination. It is often difficult or impossible to go back to the creation event after the fact to gather the information necessary to contextualize content.

The OKAPI projects of practice have demonstrated to us that the most opportune place to tag content and add meaning is when creators and curators are actively engaged with it, for the most qualified person to add the meaning also has the most incentive to make it useful, if only so that the content will be easier to find and manage.

The possibilities of what can be done with high value, rights-cleared content are boundless, and what we hope to demonstrate are that the incentives for sharing and saving are at every level. Curated information is less expensive to manage. Primary research depends on dissemination, as well as replicable results that can only be achieved if we have the original data and we trust them. Ultimately, actively cultivated and cared for content is more likely to be shareable because it is valuable to and understandable by the many communities that help to create, manage and disseminate them.

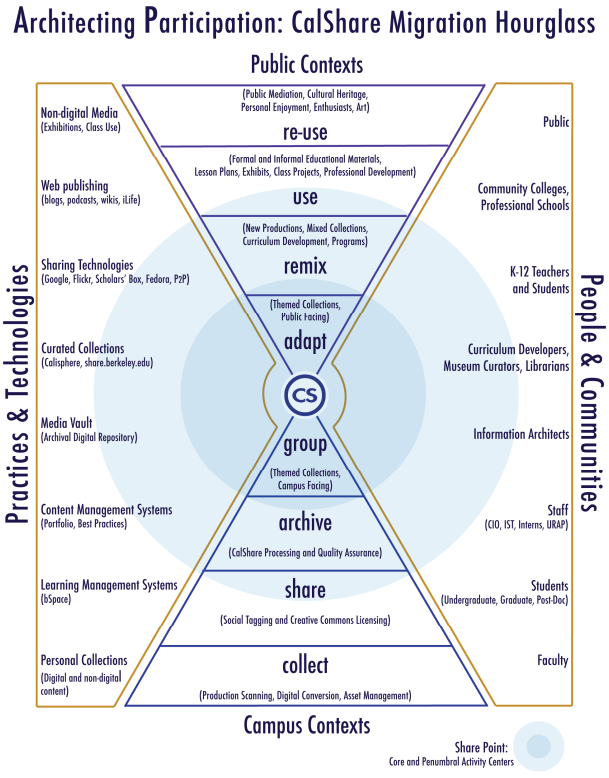

Figure 2: Architecting Participation: Calshare Migration Hourglass

We can think of digital heritage in terms of what the value is of what is being saved, its viability, how available it is to stakeholders, and how long it will last. In other words, an ideal digital heritage repository would conserve archival quality digital surrogate files in an openly accessible way, forever. This is the simplest definition of a trusted repository.

Furthermore, the Archaeology Data Service (ADS) in the UK defines the most critical factor for digital heritage sustainability is to "plan for its re-use." [AD07] Indeed, the design of decision making principles for digital heritage conservation should above all aim to the perpetual use and re-use of this content by striving to assure its reliability, authenticity and usability throughout the archival lifecycle.

We are at a unique point in history, where cultural heritage professionals must work to care for the physical past while assuring that there will be a digital record for the future. Peter Brantley, Executive Director of the Digital Library Foundation, thinks, "The problem of digital preservation is not one for future librarians, but for future archaeologists." If one imagines that the well-intentioned efforts of researchers and scholars in the modern era could be unreadable only fifty years from now, there is tremendous responsibility on individual CH professionals to insure a future for their digital work.

4. Ending on Optimism

We see the crisis not between producers and consumers of digital data, but in the capacities of cultural heritage specialists to produce the content for themselves in ways that can adhere to the principles defined by the Library of Congress and other key international standards bodies. There is a desperate need for methodologies for digital heritage conservation that are manageable and reasonable, and most importantly, can be enacted by cultural heritage professionals as essential elements of their daily work. The collaboration between cultural heritage professionals and digital specialists should lead to the democratization of technology through its widespread adoption, not the continued mystification of technology that is still being defined by the persistence of a producer/consumer divide. Born-archival content, smart algorithms that favor quality over hit count, and easy to do-it-yourself workflows are some of the keys to success.

Let me end on optimism. We can worry about the foreboding consequences of the present future, but we have the advantage of knowing the causes and potential solutions for avoiding the abyss, and the lessons of history and archaeology to guide us. We can act now, pragmatically and with enthusiasm as individuals and as a community joined in moving past the singularity so we might explain to our descendants what took place in the beginning of the new millennium.

Ideally, all of us can be carriers of the digital human genome, digital archivists in our own right. When digital file formats can provide consumers, and here we mean end-users, with digital content that is born-archival, we will have achieved the paradigm shift needed to end the reliance on digital libraries and institutions of cultural memory and potentially bring the digital dark age to a close.

We can do this every day, in little and big ways, now. This symposium is more than a little step in the right direction because it is bringing together people who want to make a difference. Let's embrace the potential of a Digital Dark Age by looking toward the Long Now and expecting that everything will be ok, and it might even be fun.

References

ANDERSON, C, 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine, [Online]. (Updated 6/23/2008), Available at: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory, [accessed 8/30/2008]

Archaeology Data Service, 2007. Digital Preservation FAQ. [Online]. (Updated 4/29/2008), Available at: http://ads.ahds.ac.uk/project/faq.html, [accessed 6/30/2008]

BAILEY, C, 2008. Institutional Repositories, Tout de Suite. [Online]., Available at: http://www.digital-scholarship.com/ts/irtoutsuite.pdf, [accessed 8/28/2008]

BARKSDALE, J, 2007. Saving Our Digital Heritage. Washington Post, [Online]. (Updated 5/16/2007), Available at: http://www.washingtonpost.com/wp-dyn/content/article/2007/05/15/AR2007051501873.html, [accessed 8/6/2008]

BRAND, S, 2003. Escaping the Digital Dark Age. Library Journal, [Online]., Available at: http://www.rense.com/general38/escap.htm, [accessed 8/1/2008]

CNN, 2007. Third undersea Internet cable cut in Mideast. [Online]. (Updated 2/1/2008), Available at: http://www.cnn.com/2008/WORLD/meast/02/01/internet.outage/?iref=hpmostpop, [accessed 8.1.2008]

GOMEZ, M et al. San Jose police: Sabotage caused phone outage in Santa Clara, Santa Cruz counties. San Jose Mercury News, [Online]. (Updated 4/9/2009), Available at http://www.mercurynews.com/topstories/ci_12106300?nclick_check=1, [accessed 4/9/2009]

GORMAN, Siobhan. Electricity Grid in U.S. Penetrated By Spies. [Online]. (Updated 4/9/2009), Available at http://online.wsj.com/article/SB123914805204099085.html, [accessed 4/9/2009]

MULVENNEY, N, 2008. IOC admits Internet censorship deal with China. Reuters. [Online]. (Updated 7/30/2008), Available at: http://yro.slashdot.org/article.pl?sid=08/07/30/1551211&from=rss, [accessed 8/27/2008]

The Long Now Foundation, 2008. About. [Online]. Available at: http://www.longnow.org/about, [accessed 6/30/2008]

WITTMAN, N, 2008. About Open Knowledge and the Public Interest (OKAPI). [Online]. Available at: http://okapi.wordpress.com/, [accessed 8/28/2008]

Figures

Figure 1: ASHLEY, M, 2008. Petabyte Age, based on Wired Magazine Illustration by Marian Bantjes. [Illustration]. (Updated 6/23/2008), Available at: http://www.wired.com/science/discoveries/magazine/16-07/pb_intro, [accessed 8/27/2008]

Figure 2: ASHLEY, M, 2006. Architecting Participation: CalShare Migration Hourglass. [Illustration].