DEEP THINKING IN SHALLOW TIME: SHARING HUMANITY’S HISTORY IN THE PETABYTE AGE

In the session, Widening The Audience: New Roles And Ways Of Producing And Sharing Knowledge For The Public

MICHAEL ASHLEY

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.

Figure 1: The Petabyte Age

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)
Danny Hillis, co-founder of the Long Now Foundation and MetaWeb, is on a mission to remind us of what matters. He is an expert in digital preservation and is a good a source as any to get a sense of how things are and how they might be. He is an optimist, generally speaking, but he has doubts about the state of our cultural memories if we rely on corporations to define time and priority for us. Like Hillis, I am an optimist, but in the domain of digital preservation, we must ‘hope for the best and plan for the worst.’

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 of this year, 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. The Internet, and digital technology, remains volatile, friable and at high risk from the perspective of long-term human history.

Sometimes access is cut by accident, but more and more often it is intentional, targeted and disturbing. Figure x shows a map of locations in the world where access to content is being aggressively denied. On September 29th of last year, the Burmese military Junta shut off the Internet completely for the country to quash freedom of the press (Global Voices Online 2008). This month, the 2008 Olympics in Beijing were tarnished by deals cut between the International Olympics Committee (IOC) and the Chinese government to block sites that were “deemed sensitive to its communist leadership” (Mulvenney 2007). The Internet is not free; it is a commodity of power.

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).

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.

Do we leave these efforts to Google, Microsoft, Apple, Yahoo, should we? Will computational semantics really end scientific inquiry, and can our Web 2.0 darlings of innovation be trusted as the stewards of human history? 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.

**Resetting Our Clocks To Multi-Generational Time**

The Long Now Foundation, established in “01996” seeks to “become the seed of a very long-term cultural institution,” meaning adopting a counterpoint to today’s “faster/cheaper” mind set and promote “slower/better” thinking (Long Now Foundation 2008). Archaeologists as well as archivists don’t think that a decade, a century, or even a millennium is a long time when we take a look at it from the perspective of the human record.

Let’s take a look at “now”. Today is now, we can think of today as now, that is. We can stretch this definition to be more ‘nowish’, yesterday and tomorrow perhaps if we want to think of now as a place in time with a beginning, middle and end. Nowadays, personal time, perhaps a decade, the millennium, the 90’s, the roaring 20’s, the beginning or end of a century, half a lifetime these days, the overlap of generations past and forthcoming, history. Archaeologists are comfortable discussing and pondering a span of 10,000 years past - if we extend this time definition forward, we have a “now” that has been recalibrated to human time.

Looking back, then, at the last 10,000 years, we have made indelible impressions on our planet, but we can also be reminded that our current now has only just begun. The digital revolution is a remarkable experience and achievement in technology, no less so than the great pyramids or the construction of the first cities, but will it last, and how will it transform in the next hundred years or so?

If we look at the Long Now as a history of communication, we can look back much deeper and remember there was a time before human speech. A time where we inscribed our thoughts on cave walls (30,000 years before now), and we can still debate whether this is “art” or not. At 10-15,000 years ago, an explosion of symbolic expression worldwide.

**Figure 3: The Long Now of Human Action and Communication**
Within our current Long Now, we can look at remarkable achievements in personal and social communication, from the wall paintings at Çatalhöyük, to Sumerian cuneiform tablets and Egyptian papyrus. Drums and smoke signals were ways of reaching people across long distances. Expressed more person-to-person, private networked messaging is made possible through the early postal systems that come on the scene as far back as 2200BC.

The heliograph, devised around 450BC, provided much more accurate networked messaging through ‘morse code’ by mirror reflection. Julius Caesar got the word out through the first newspaper around 43BC, and yet paper is really a ‘now’ technology, 100CE, some 3000 years after cuneiform.

And suddenly, relatively speaking, the tools needed for personal communications are developed in the very recent now with the invention of the pen and paper. Just yesterday, then, the capability for precise, coherent, personal, private, networked, social, scalable communication between people is here. Telecommunications, electronically based, are technologies just 150 years in use. And computers as we know and use them today really won’t reach development and adoption until much sooner, say the 1960’s.

Some critical developments, for which we would not have an Internet, are what we may call singularities, technological advances that have transformed society. In 1879, the electric light is successfully tested by Thomas Edison, and a year later he forms the first electric power utility, offering electricity to power cities, businesses, and homes. In late December of 1941, nuclear power is under the control of man, and the capability of erasing the human race is born. This is a fateful moment, because this capability is a prime motivational factor for the creation of the Internet, to provide redundancy for the US military in case of nuclear war.

The Internet, a singularity for human communication, the World Wide Web, is realized just 25 years ago in 1983. Can you imagine our world without it? Some of us can remember 1978 when the Internet was Star Trek, science fiction. Yet the Internet, or electronic networking, is transforming so quickly, what will it be like in 25 years, one reflection of time from its creation?

**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. Examples include developing the campus-wide Media Vault repository system, creating new models for sharing digital collections as part of the U.S. Department of Education Scholar’s Box Project, and building a virtual humanities laboratory as a demonstration project for the international Project Bamboo consortium. Recently, OKAPI is working within the emerging research domain of synthetic biology, engaging diverse academic and public communities in reflection and dialogue on how economic, political, and cultural forces may shape the development of synthetic biology and on how synthetic biology may impact human security, health, and welfare.

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.
We produce more content now than it is humanly possible to preserve. Current estimates are that in 2006, 161 billion trillion bytes - 161 exabytes, or 161 billion gigabytes - of digital data were generated in the world -- equivalent to 12 stacks of books reaching from the Earth to the sun. In just 15 minutes, the world produces an amount of data equal to all the information held at the Library of Congress (Barksdale 2007).

Hillis describes the ‘here and now’ as a ‘Digital Dark Age,’ because information is devalued by the ubiquity of digital content that cannot outlast our lifetimes (Brand 2003). While we have more than enough storage media to hold the cultural memory of the planet, the half-life of data is currently about five years. This is due to the fact that digital preservation is not a corporate priority, nor a consumer priority at present. It must, therefore, be a producer priority.

Digital archivists resist new file formats, new metadata standards, new lifecycles and practices for all the right reasons. Consider the fiduciary responsibility of institutional repositories that are charged with keeping content safe, archival, accessible, for as long as possible. Minimizing file formats and standardizing metadata minimizes risk (and presumably, costs) as formats become obsolete. The problem is that by limiting the formats archives are willing to accept, we are actually putting the great majority of digital knowledge at risk. JPEG and MP3 are just two examples of ‘lossy’ file formats that are ubiquitous and also not acceptable by most ‘trusted’ repositories. Is the information within these files meaningless?

If we wish to avoid a Digital Dark Age, we need to compel consumers into action. In this case, the consumers are the archives and the algorithm, the bankers and indexers of our data. To do so, there are several strategies we can apply. We suggest that we need to design digital media to be ‘self-archiving,’ adaptable to virtually any digital environment so that there is no need to rely on ‘institutional’ repositories to exist, at least not in the monolithic sense. We need file formats that are too clever to ignore,
that minimize risk while maximizing semantic meaningfulness, and can transmogrify without degrading as they move across the cloud.

John Kunze, preservation specialist for the California Digital Library, calls for ‘born-archival’ media that is fully accessible and preservable at every stage, throughout the lifecycle of this data, from birth through pre-release to publication to revision to relative dis-use and later resurgence. Data that is born-archival can remain long-term viable at significantly reduced preservation cost (Kunze 2008). Several organizations – Cultural Heritage Imaging, UC Berkeley Media Vault Program, ICS-Forth in Greece, UC Santa Cruz, CNR-Pisa, Minho University in Portugal – are working on developing the services, semantic wrappers and file formats that make born-archival media a reality.

We need institutional repositories to exist. Clifford Lynch, Director of the Coalition of Networked Information, reminds us that institutional repositories are “most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.” (Bailey 2008). Until we can invent the digital equivalent of cuneiform tablets, that is, a substance that can preserve the medium and the message equally, we will need stewards of the human record. Our short-term proposition (for the next few decades, say), is to provide digital archives a revolutionary way forward in sustainability through semantically managed archive ready technology.

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.

**Empowering Individual Action Through Institutional Collaboration**

We advocate for both individual professional responsibility and multi-institutional, multi-disciplinary curatorial management of digital heritage content for the foreseeable future. Unlike the physical archives of the Library of Alexandria, lost forever to humanity, digital heritage can be in more than one place at a time and in more than one form, potentially assuring its longevity despite the ephemeral nature of the media. This multiplicity of location and form is both the promise and the peril of digital heritage.

With increasingly diverse data formats, larger file sizes, changing media types, distributed databases, networked information and transitive metadata standards, how are today’s heritage specialists to plan for such an uncertain virtual future? It is increasingly difficult for individual scholars and researchers to do the right thing when it comes to digital heritage conservation. The accountability for the conservation of digital heritage falls to all in the CH field, but what is a reasonable course of action in the face of such adversity?

The importance of developing sensible plans to preserve our digital heritage cannot be minimized. Responsible preservation of our most valued digital data requires answers to key questions: Which data should we keep and how should we keep it? By digital heritage conservation, we mean the decision-making criteria to discern what must be saved from what can be lost. Everything can't be saved nor is it desirable to do so. How is this data to be saved to ensure access in five years, 100 years or 1,000 years? In the next 100 years, we will go through dozens of generations of computers and storage media, and our digital data will need to be transferred from one generation to the next, by someone we trust to do it. Finally, who will pay for all this?

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.

**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.

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


Figures


Figure 3: Ashley, M, 2008. *The Long Now of Human Action and Communication*. [Illustration].

Figure 4: Ashley, M, 2006. *Architecting Participation: CalShare Migration Hourglass*. [Illustration].