PREPARING FOR THE AGE OF THE DIGITAL PALIMPSEST

Jason Bengtson, MLIS, AHIP

KEYWORDS: IT, informatics, preservation, digital forensics, data curation, preservation,

palimpsest

Author Bio: Jason Bengtson is the Emerging Technologies/R&D Librarian for the University of New Mexico's Health Sciences Library and Informatics Center in Albuquerque, NM. He is a graduate of the University of Iowa School of Library and Information Science in Iowa City, Iowa.

INTRODUCTION

We, as a society, must make our choices.

This paper is about one of those choices, the choice of what we preserve and the repercussions of that choice for the field of Information Science. It describes the modern phenomenon of deleted and partially overwritten digital data and positions it within the larger historical context of data partially overwritten in previous eras and in previous formats. At its core, this paper's argument is simple; just as Information Science sub-fields of various types have had a large part in the work of discovering and recovering such overwritten, or *palimpsested* texts in the past, Information Science should take a leadership role in the recovery of these new, digital palimpsests. This issue is not simply one of technology. It is an issue of the management and recovery of information; an issue that includes technological considerations but also transcends them. Digital palimpsests will require the efforts of many disciplines, including Information Technology, Preservation, Archival Sciences, and Archeology. But it is Information Science that stands at the crux of how these other disciplines must approach this crucial issue, and it is on Information Science that the burden of managing these competing interests should fall.

THE SIGNIFICANCE OF PALIMPSESTS

It is not always obvious to those outside of academia that preservation takes work and resources. It requires the will to set aside those resources over time,

sometimes in perpetuity. Because of this, the simple fact is that society can't preserve everything. In fact, given the very finite nature of human resources, we usually can't even preserve everything of a given type. Can we preserve copies of all the works by Ernest Hemmingway? Probably, if we're careful (which is to say we can preserve copies of all the works he presented to the world, or that we recovered after his death). Preserve copies of all literature? As works were lost to history that became impossible long ago. Even preserving copies of all literature of the twentieth century would probably be impossible, especially if a less essentialist definition of what constitutes literature were used as a guideline.

Because we can't preserve everything, we as a society and we who are in the information business must make choices about how to allocate our preservation resources. We must also live with the fact that, as time passes, some of these choices must be revised.

These facts were not lost on those who came before us. In the middle ages, while the cost of a scribe was paltry, the cost of parchment and binding was enormous. Some of the largest medieval libraries in Europe had only a hundred or so volumes. Never was it clearer to the caretakers of knowledge that not everything could (or should) be saved. And so we see today that many parchment leaves were erased and reused to record something else.

This type of erased document is known as a *palimpsest*. Literally meaning "scraped again" the term refers to the initial act of scraping an animal skin to create parchment, followed by a re-scraping (and often acid treatment) to erase the first document so that the expensive leaves could be reused.[1,2] This was done for many

reasons: a text could become illegible due to changes in linguistic currency, a particular text might contain laws that had become void, or the text might consist of outdated liturgical material.[2]

Whatever the reasons behind it, palimpsesting became a relatively common practice. So common that a tremendous amount of scholarship has gone into trying to uncover the original texts of medieval palimpsests. Nowhere is this more evident than in the extraordinary work that William Noel[3] and his team of experts have conducted on the Archimedes palimpsest. Perhaps the greatest scientist of antiquity, Archimedes of Syracuse was renowned for his inventions and scientific work, but our sources from antiquity about this genius are limited. Using state of the art imaging and pioneering paleographic forensics, Noel's[3] team has uncovered much of the work of Archimedes recorded in a palimpsest purchased at auction by an anonymous benefactor in 1998, including exciting new discoveries about the sophisticated nature of Archimedes' mathematical achievements.[2,3]

ENTER THE DIGITAL PALIMPSESTS

The saying goes that everything old is new again, and so it is with the palimpsest. Today it is not the second scraping of animal skins but the deleting of digital files that promises to create a new field of scholarship for the future Information Scientist. As digital archives have proliferated, only to die a quiet death, their storage space has been reused. Hard drives that once archived masses of forms and digital images (to name only two of a dizzying myriad of digital archive types) are now home to

shared intranets or a new generation of documents. The files are deleted and their storage mediums reused, and then the data is often forgotten. But, as we've seen with the palimpsests of ages past, they have not necessarily been lost. Modern forensic techniques, most developed for law enforcement, can recover a surprising amount of "deleted" data. Files can also be recovered from unallocated disk space or from disks that lack a file system through the use of file carving processes.[4] In the future we may even see technologies that can recover some overwritten digital data . . . a possibility that for now remains out of reach.[5] The information lies dormant, slowly degrading, as in the third codex of Archimedes, waiting for the day when it will be rediscovered, and patiently restored.

So, as with physical palimpsests, the inevitability of digital palimpsests is made apparent because of two facts, the second being a consequence of the first. One, we have to make choices about what to preserve, and two, we will decide at some point in the future that some of those choices were poor ones.

We have already been faced with this situation. The Apollo eleven moon landing provides us with an excellent example. The original video tapes of the moon landing probably still exist somewhere, but NASA was unable to locate them after an exhaustive search.[6] The agency now believes that they were erased and taped over years ago in response to a shortage of high quality magnetic tape caused by a lowering of manufacturing standards in the industry.[6] The original material by NASA simply no longer exists . . . a victim of the sudden increase in value of the medium upon which this visual text was stored. The tapes would have been degaussed (erased by removing the magnetic signatures through which the information was stored) before being

overwritten.[6] It is unfortunate that these tapes were probably degaussed and overwritten. Given the limits of our current technology, even if the tapes were found it is unlikely that much of the moon landing text could be recovered from them. Fortunately, despite the probable erasure and loss of these palimpsested texts, NASA was able to locate a number of lower quality versions of the broadcast which were then reviewed for selection based on comparative quality. The best videos (some fifteen scenes) were restored and enhanced by Lowry Digital, a company known for restoring old movies.[6]

These are not unusual problems at NASA. A report from 1990 by the General Accounting Office detailed to an excruciating degree the dangers faced at the time of their investigation by space data stored at the agency on magnetic tape.[7] These included lack of back-up systems, lack of security, and environmental conditions that could enhance degradation of the materials.[7] If this is, or even was, the state of affairs at one of the nation's most technologically imbued agencies as it safeguarded national information treasures, what can Information Scientists hope for from the countless number of other individuals and organizations that lack any sense of the potential temporal importance of the information they record? How much research data is not being warehoused and safeguarded properly by researchers at major Universities? How much business data is being lost? Still, raw data and its scientific value are not the primary concern of this paper. The value of such information in a cultural and historical context, a consideration that will likely outlive the scientific value of the data, is at the crux of the issue of digital palimpsests. Later in this paper the author will attempt to define a new paradigm for the conceptualization of such palimpsested contents, but first it may be useful to examine some aspects of data recovery as it is found today.

THE CURRENT STATE OF DATA RECOVERY

There are two main strains of data recovery currently in existence that have a direct bearing on digital palimpsests. The first is the field of Digital Forensics. Digital forensics is a field based around the needs of the legal profession, both in criminal and civil venues, as it pertains to the recovery of data. Digital forensics involves recovering data in a way that is as free of distortion and bias as possible, up to and possibly including providing testimony on that data in court.[8] Due to the amount of data recovered, there is often a heavy analysis component to this work, along with increasingly serious requirements for the examination of networks, rather than just the examination of individual devices.[8] While increasingly sophisticated tools have been developed to recover data from portable media, like flash drives, or to recover data from drives without the use of a nominal file system, these tools are largely designed around the needs of the legal system; particularly law enforcement[9,4]. Those unique needs that form a part of cultural and historical recovery efforts, which will be discussed later in this paper, are simply not part of the problem that digital forensics must grapple with. Nonetheless, arguably the most dynamic work in the area of recovering deleted data is to be found in this field, making it vital that those tools and techniques be utilized (albeit in modified forms) as part of the technical side of recovering the texts of digital palimpsests.

A very different approach to data recovery can be found in the field of Digital Archeology, which has been pursued with great vigor in the United Kingdom. Here can be found a discipline that is concerned with digital artifacts as cultural objects. The work

of individuals such as Seamus Ross and Ann Gow[10], as exemplified by their report on the subject to the University of Glasgow's Humanities Advanced Technology & Information Institute, has helped define digital archeology. Projects like the recovery of the data from the Danebury hill fort excavations have been key contributions of digital archeology to the humanities.[11] However, as the report for HATII demonstrates, the interests of digital archeology remain largely confined to questions of degradation or obsolescence of the storage medium or software technologies.[10] The work done on the Danebury fort excavations, for instance, is perhaps better characterized as data migration than true data recovery.[11] The data digital archeology is rescuing is data that has been preserved, but preserved poorly, rather than data that has been deleted or partially overwritten, as is the primary concern of this paper.

Clearly both of these disciplines have a great deal to offer to the recovery of digital palimpsests, yet neither of them, at present, is an adequate paradigm on its own. Principles of both fields must be combined in order to adequately recover, study and preserve the digital palimpsests to come. Yet, even when these two disciplines are combined, there may be room for additional lessons from the study of physical palimpsests from antiquity.

DIGITAL PALIMPSESTS AND THE ROLE OF INFORMATION SCIENCE

It is reasonable to ask why any of this is an Information Science problem. It has been shown already that much has been done in the area of data recovery by Information Technology, in the form of both digital forensics and digital archeology. In

the realm of digital archeology the humanities have become involved as well. But, as with so many things, while IT can provide the tools, it is the Information Scientist who can best employ them. Both the humanities and the hard sciences have long relied on librarians to take the lead in adding value to information. As the field that has traditionally managed the storage and cataloging of information, and as a field that has been inextricably tied to studies of physical texts, Information Science is the logical choice to take the lead in recovering the texts of this new, digital landscape. Digital palimpsests will bring together cultural and technical considerations in a way rarely seen in disciplines other than Information Science. And Information Science, with its history of managing multiple disciplines in the management of information, is uniquely positioned to provide both the vision and the oversight needed for digital palimpsest recovery to be as successful as it needs to be.

In recovering and investigating future digital palimpsests, it will be the work of the Information Scientist to find, then make sense of, information that had once been lost. It is the duty of the Information Scientist to use many tools together (information retrieval, physical conservation, context research, data recovery, data conservation, data migration, and reformatting) to retrieve what can be restored in the digital palimpsests that will fill the future.

Just as texts were palimpsested for many reasons in the middle ages, they are palimpsested for many reasons today. Their value isn't understood, the storage space they take up is seen as more valuable than they are, or our priorities as a society simply change. It isn't so hard to imagine the work of a modern day Archimedes being deleted to make way for digital music, or copies of a reality television show. Every day we are

making poor choices in both short term and long term digital conservation. It is inevitable that those decisions will come back to haunt us. When they do, the Information Science field must be ready to face this challenge. We must prepare for the palimpsests to come.

A CHANGE IN PERSPECTIVE

Fundamental to the challenge of digital palimpsests is the need to view them as true palimpsests. To put it another way, we must begin to see these objects of recovery as texts rather than as data. A text, in the Literary Studies and Library Science senses, can be anything from a work of literature to a grocery list. The defining quality of a text is not its format, but rather its position as a discrete arrangement of signs. Signs can have multiple meanings, based as much on the audience as the author, and their relationship to the signified (the word apple in relation to a physical apple, for instance) potentially gives them a dimension that units of data lack.

Moreover, a text is an arrangement of signs, yet it is a discrete entity that relies on the interplay of its constituent signs to create it. It is a discrete entity that is a function, not only of its constituent parts, but of the interplay between those parts, making the whole more than their sum. This gives texts a fundamentally different role from that of collections of data, which, while of vital importance to Information Technology, often do not possess the same implication of interdependent cohesiveness, completeness, and interpretive richness as a text. Recovering a digital palimpsest is not about recovering data except on the most technical level. In a larger

sense, the effort is about recovering a text, something complete in and of itself, that becomes a primary source in a larger canon. To say that seventy percent of a text was recovered is to say something very different than seventy percent of available data was recovered. The data lost may have been more or less relevant in relation to the data which was recovered. The data may have been in several discrete data sets (themselves a little like a text) which were more or less complete. But a text is something that exists as a recognizable whole. It is a primary source to be “read” and analyzed on many levels, and which can, therefore, communicate data on many levels. This fundamental, semantic difference must form the basis for future efforts at palimpsest recovery.

SHAPING THE TOOLS

As mentioned previously, most digital recovery of deleted data is centered upon the needs of law enforcement, not paleography (the study of ancient writings). Information Science, as a profession, needs to work to create a second point of focus for these efforts. Information Science needs to bring the principles of conservation and paleographic study to the forefront of the development of data recovery tools. Information Science needs tools that recover data nondestructively, preserving the original digital information in the same way that conservators strive to preserve the physical document of palimpsests, so further discovery can be done as data recovery technology improves. Information Science must never forget the earliest work that was done on palimpsests using chemical reagents that irreparably damaged the physical

manuscripts, limiting the lifespan of the manuscript and making the later use of more modern tools difficult to impossible.[23] As highly as we may think of our tools for data recovery today, history has taught us that they will be eclipsed by the tools of tomorrow, including some that may incorporate principles we could only imagine. These tools are useless, however, if the original media is lost or further damaged by crude attempts at recovery.

We should also remember our other principles of physical conservation. The palimpsested data may be important, but the material that replaces it may also have scholarly value and so should not be discarded or destroyed out of hand.[3] The modern scholar delights at the discovery of ancient texts, even when they prove to be nothing more than bills of sale. The study of charters and other documents (the discipline of Diplomatics), for instance, is a vital part of Medieval Studies.[2] The great texts may garner more attention, but it is often the everyday detritus of humanity that has given researchers the most useful information about the past. Data recovery tools should seek to recover all of the palimpsest texts on a storage device.

LESSONS FROM ARCHEOLOGY

Some of the most useful finds in archeology have been from garbage heaps. So it will be in the future, as the discarded materials of this age become the recovered artifacts of the next. Information Scientists, more than most, should take care to remember that the garbage of the past is often the meat of scholarship. Librarians must be vigilant, on the lookout even now for signs that an object requires further study. What

if a librarian or an archivist had been there on the day that a bank of old videotapes was pulled for degaussing at NASA? Perhaps a better question is, would an information scientist today know what to do if they found an old hard drive at the yard sale of a JPL employee, or the obsolete desktop computers of the local historical society (complete with a deleted copy of their ill-fated archive of scanned records) in a dumpster? Would the librarian know enough to preserve the storage media, and protect what might be locked inside? By laying the groundwork for this discussion now, the librarian of the near future may be more aware of these issues. By developing this awareness now, Information Science can construct a future where valuable data can routinely be rescued from digital storage devices that otherwise would have simply ended their lives under a pile of refuse.

ADDING VALUE

Arguably, the most defining role of librarians comes through adding value to information. Librarians make information worth more than it otherwise would be, and they do this in two main ways: first, they make information available (through collection development, the science of cataloging, digitization, or through search and discovery) and second, they provide context for information. They give information a place within a larger world that provides it with greater meaning. Take, as an example, the statement, “the day became as dark as night”. This phrase has limited intrinsic value. What does it mean? Maybe it refers an eclipse, or a storm. Perhaps it is just a literary flourish. Maybe it describes the day literally transitioning into night. Without context, this information has

limited useful value. But if the previous statement was substituted with, “because of the Krakatoa eruption, the day became as dark as night” the information becomes more meaningful through the simple addition of context. This is a basic example that runs the risk of making the process of context placement and addition seem easy. Of course it is not. It requires research, and scholarly effort of a type not always appreciated outside of academia. Yet it is often by finding and then meaningfully vetting information that Information Science drives the other sciences.

The same must be done for digital palimpsests. Their origins must be explored, cross-checked and validated. Individual pieces of metadata must be connected together, producing clues that will allow librarians to give the texts within the storage device being investigated a contextual setting of time, place, ownership, condition, etc.

MANAGING THE EFFORT

When reading the Archimedes Codex, it is easy to get the impression that William Noel felt like he was in a little over his head with the Achimedes palimpsest. He readily admits in the book that he co-authored, that he felt the project was outside his area of expertise.[3] But Noel is a curator and curators, like librarians, know how to draw together information and resources. His work has been extraordinary, despite his lack of knowledge in those areas needed for the palimpsest restoration. Noel sought out the proper expertise and brought the necessary resources together.[3] Other Information Scientists need to be prepared to do this as they confront digital palimpsests. They must

seek out the knowledge and the expertise from their fellow scholars in many disciplines as they lead the effort to investigate such texts. Nonetheless, even as they draw upon that expertise they should continue to position themselves centrally to the effort. Digital palimpsest recovery will be about the management of information; all other facets of the problem are secondary. Librarians are the leaders in the field of information and they should not invite others to take their place.

CONCLUSION

As librarians we are Information Scientists, which makes it our duty to look to the future of information. Information Scientists must prepare vigilantly for the new challenges to our profession that await us. Among these challenges will be digital palimpsests. We must combine the strengths of the fields of digital forensics and digital archeology with the hard-learned lessons of our own discipline in order to have the tools and the methodologies in place to confront the unique considerations posed by digitally palimpsested texts. Information Science must be ready, not because there is no one else, but because there is no one else as well equipped to face the special needs of this challenge. Rather than seeing this as a burden, librarians should look to the digital palimpsest as an opportunity to further refine and define their academic field. In many ways, especially given the changes wrought by digital technology, Information Science is still a young discipline. Like all academic disciplines, Information Science must work to define itself and establish its utility if it hopes to attract talent, scholarship, and funding. By positioning itself at the forefront of a concept such as this one, rather than

attempting to react to it once it presents itself, Information Science can continue to become a more refined, defined, and recognizable academic discipline.

1. Oxford ED. "Palimpsest, n. and adj.".

2. Clemens R, Graham T. *Introduction to manuscript studies.* Ithaca: Cornell University Press; 2007:301. http://www.loc.gov/catdir/toc/ecip0714/2007010667.html.

3. Netz R, Noel W. *Archimedes codex : How a medieval prayer book is revealing the true genius of antiquity's greatest scientist.* Boulder, CO, USA: Da Capo Press; 2009.

4. Yoo B, Park J, Lim S, Bang J, Lee S. A study on multimedia file carving method *Multimedia Tools Appl*. 2011.

5. Wright C, Kleiman D, Sundhar R.S. S. Overwriting hard drive data: The great wiping controversy. In: Sekar R, Pujari A, eds. *Information systems security.* Vol 5352. Springer Berlin / Heidelberg; 2008:243-257. http://dx.doi.org/10.1007/978-3-540-89862-7_21.

6. NASA. The apollo 11 telemetry data recordings: A final report. . 2009(9/17/2011).

7. U.S. GAO - space operations: NASA is not properly safeguarding valuable data from past missions . ;2011(9/17/2011).

8. Cardwell K, Books24x7 I. The best damn cybercrime and digital forensics book period. . 2007.

9. Breeuwsma M, de Jongh M, Klaver C, van der Knijff R, Roeloffs M. Forensic data recovery from flash memory , high tech *SMALL SCALE DIGITAL DEVICE FORENSICS JOURNAL*. 2007;1(1).

10. Ross S, Gow A. Digital archaeology: Rescuing neglected and damaged data resources. . 1999(9/17/2011).

11. Digital archaeology rescues lost records for danebury hill fort: Oxford ArchDigital restores early digital data. *Records Management Bulletin*. 2003(117):18-18.