The Digital Vapor Trail:

Why Early Digital Assets Merit Special Attention

Chris Muller, Muller Media Conversions www.mullermedia.com

Editor Note: The following are a series of vignettes from the author's experiences in data-rescue projects

Let's picture a digital technology jet plane racing along, leaving a vapor trail of incompatible, deteriorating media, undocumented files or file-systems, myriad backup tape and optical disk formats -- it's not a good thing.

Some organizations possessing "at risk data" have created special tools and performed great "data-rescue" projects. Even better, they were scientists and researchers who then proceeded to analyze, publish, re-purpose, etc. But often, the tools for rescuing older digital assets are not readily available to the custodians. Our theme here is that, just as digital data can get lost in the vapor trail, some important *data rescue capabilities* are also fading away. Later, the author will share some thoughts about creating a non-profit to preserve and manage data rescue capabilities.

A Great Pre-Digital Data Recovery Example

12,000 pages of New Amsterdam records sat largely un-noticed in an Albany vault for over 300 years and were then re-discovered in the late 1960's. They were lucky enough to find a great scholar expert in 17th century Dutch language and writing style—which we've learned is totally unreadable to modern Dutch speakers. Ongoing translation efforts lead to Russell Shorto's wonderful book "*The Island at the Center of the World*". Luckily, those documents had the "*Luxury of Languishing*". However, with pre-digital information:

- That which remains was recorded on lasting material.
- We can often *see* what's needed to be read, translated or copied.
- Optical scanning, OCR and digital photography are mature and ever-improving tools

But, regarding older digital material:

- If only older magnetic media would last as long as those New Amsterdam records.
- If only we could read arcane media/files without special equipment and software.
- If only some of the tools and skills to deal with those weren't also fading away.

Here's a view of the obstacles to recovery and conversion of legacy data; a bit like peeling back the layers of an onion:

Media Compatibility
Age and Storage Conditions
Recording Method
- -*
Operating System/Filing System
Backup, Exchange, or Archiving Software
Application File Structure
Application File Encoding

*Virtual tape and disk copies can often protect the bits and bytes, allowing elbow-room to tackle the other layers as time and funding permit.

While there are challenges with old media and arcane files, there are also some advantages. Back in the day, computer media was expensive and hard to use—with little in the way of tweets, mp3s, politics and porn. So the "value density" is often greater than that on much of today's media. Also, a good ratio exists between older media capacity and inexpensive new media.[1]

Data-Rescue and Conversion Can Be Fun!

Learning to work with and continue to improve existing equipment and software tools; puzzling out early, sometimes unique data formats; these challenges can be very rewarding. Even better, you get to work with scientists, researchers, historians and archivists who really understand the underlying value of the data. Here are a few examples that I hope you'll find interesting from our *"Tales from the Digital Crypt".*

Cryptie #1: Some pre-Whitewater Fun

Think "Clinton and Moscow". We're talking about a recent Secretary of State visiting Russia, right? Nope. In 1992, Mr. Clinton was a contender for a presidential nomination. There were rumors that as a college student in 1969, he had gone there to protest the Vietnam War. Political rivals got curious and issued a FOIA request. An old State Department tape came to the US Attorney's office in New York. Guess what? There was no metadata (that is, no file format info) --a sadly common occurrence. But the D.A.'s office found some techies who enjoyed hacking away, so that the plaintiffs could receive useful, readable information. This stands out in our memory because of a few other things that happened around the same time…

---

[1] One example: the content of 9,000 government mainframe cartridges fit on one inexpensive USB hard drive, so future backup/migration efforts are tiny compared to reading all those old tapes.

ASSOCIATED PRESS JANUARY 3, 1991 WASHINGTON--A slice of America's history has become as unreadable as Egyptian hieroglyphics before the discovery of the Rosetta stone. Vast untold volumes of historic, scientific and business data are in danger of dissolving into a meaningless jumble of letters, numbers and computer symbols. Much information from the last 30 years is stranded on computer tape from primitive or discarded systems-unintelligible or soon to be so…

Due to this article, along with radio commentaries by Charles Osgood and others, the issue of "data at risk" got a lot of attention. One of the people mentioned was the iconic Dr. Ken Thibodeau of the National Archives and Records Administration. —At my wife's urging I contacted Dr. Thibodeau. Although we had many years of experience in media conversion, we had little exposure to government records. Unfortunately, we did not hear back from NARA for over a year.

In the meantime, there was revived Congressional support for digital preservation. It also turned out our new president (Mr. Clinton) was a strong supporter of the NARA and records preservation[2] and approved new funding support.

Helping NARA to build that Rosetta Stone

If you enjoy history, you can imagine how exciting it would have been to be interviewed in

FDR's "fireside chat" room. In fact, it was nowhere near a fireplace! President Roosevelt had used a small studio (now a conference room) in the depths of the NARA building on Pennsylvania Avenue. Subsequently, NARA issued a contract for the development and support of something called "*APS*", the Archival Preservation System, beginning a happy 14-year relationship between a small company in New York and great digital preservationists in DC.

Cryptie #2: Some Watergate Fun

In the early 1970's, a mainframe computer was used to store President Nixon's appointment calendar and notes. 25 years later, the only copy of the information was contained on an old tape reel. Historical researchers wanted access to the information, but they had two problems: (1) the tape was in danger of decay, and (2) the data format had never been identified. They asked their techies to convert the data and create a new program for researchers. Luckily, the analyst had done work with Vietnam military records and noticed similarities in the data structure (from a program known as NIPS). Of course, we all hoped there would be Watergate secrets revealed, which did not prove true. But it was interesting to baseball fans like us to see that the first visitor to the Nixon Whitehouse was Hank Aaron.

[2] Just don't ask about Sandy Berger's visits to NARA.

Cryptie #3: More Whitewater Fun

In the 1980's an attorney for the Rose Law firm in Little Rock, Arkansas did legal work for a real estate project named "Castle Grande". By early 1994, all related paper and computer records had mysteriously vanished (or were "vacuumed[3]"). The firm requested data recovery. Obviously, the perpetrator was not a big fan of digital preservation—but he did not realize that those disks are not so easily erased.

IPUMS and Crypties from Around the World

IPUMS-International[4] collects, analyzes, privatizes and publishes population data from around the world. Tapes and disks arriving from many countries, such as Bangladesh Egypt, Kenya, Mali, Mexico, Nepal, Pakistan, Peru, Romania, Santo Domingo, Sudan and Turkey. One of our most interesting projects involved six weeks in Dhaka, at the Bangladesh Bureau of Statistics. Very capable people, they had been confronted with daily power outages and other problems making it impossible to store legacy tapes in an optimal way. Many of their tapes suffered from decomposition. A system was installed to read and convert the files, plus tape cleaners and training for BBS Staff.

One lesson learned through our work with IPUMS: *there is a world-wide need for Data Rescue.* But many nations do not have the financial resources to put such projects at high priority. That's why groups like IPUMS with funding from NSF and others often reach out to help. In other cases, collections of "at risk data" have yet to be officially noticed.

Old Professors' Stashes (*OPS*)

"I'm ready to retire and those old data tapes are my legacy!" The professor or his colleagues take a new look at his data and realize that the old health data, combined with population and climate information may well produce remarkable new insights. Re-discovered stashes like this are cropping up all the time! Of course they're not always from an old professor, but it's a fun way to characterize valuable scientific or historical data that, at least for a while, have been overlooked. Here's a quick look at three data recovery cases:

[3] A hand-written note from a Whitehouse staff meeting several weeks before the records disappeared contained the phrase "*vacuum Rose records*" (apparently a total coincidence).

[4] Part of the Minnesota Population Center.

Physical Science Case

This data was from a forestry genetics project that began some 35 years ago. The intent was to study various timber species and see how they would do if planted at different altitudes, longitudes, and moisture-levels. Later, a key player realized that the details gathered by the original study could be re-purposed to gain insight into "tree migration", a process that normally takes centuries. This could enable researchers to know which species will do best under certain aspects of climate change, and even to assist the migration process. Floppies were from older Apple drives[5] in a variety of formats, but the real challenge was that the diskettes were "flippy floppies", with each side written as a separate volume. Currently available floppy drives require index holes. And when a diskette is flipped over, the hole is no longer in the right spot. To do the flip side we had to punch extra holes in the diskette casing without scratching the floppy. The "*highly sophisticated"* tool we designed is pictured to the right.

Social Science Case

Like the fireside chat room experience, picking up well-packed census tapes at a Consulate certainly adds some flair to a project. The micro-data had been backed up to tape in what on the surface seemed to be ANSI standard format. But somehow, a type of compression had been applied to the individual records. Similar to the Nixon tape experience, we were lucky enough[6] to have seen that compression before.

Cultural History Case

Dr. Frank Siebert dedicated much of his life to safeguarding the Penobscot Native American language. After many interviews with native-speakers he created a full dictionary using special software, stored on diskettes that were unfortunately not at all compatible with standard file systems. Character encoding was also morphed for unique display hardware. We had the honor to work with the American Philosophical Society and Maine Folk Life Center on the project to ensure the preservation of this legacy treasure.

Thoughts on Preserving Data Rescue Capabilities

There are some great Data Rescue organizations addressing specific needs. One example: the International Environmental Data Rescue Organization (IEDRO), focused on recovering older environmental observation data. They have leadership with the know-how to evaluate the scientific value and also perform the physical recovery and digitization.

Other groups have focused on certain things such as cataloguing the myriad PC file types and in some cases creating open source conversion software. Another example, the CODATA Data-at-

[5] Those particular drives didn't rely on track index holes.

[6] Can't resist this quote. A young Lee Trevino was asked if he felt lucky to have gone from a poor Latino teenage caddy to a world-famous golf pro. "Yup," he said "and the more I practice the luckier I get."

Risk Task Group, has several projects, one being creation of an inventory of important scientific data at risk of loss. And of course, there are computer museums to preserve old systems—some working, some not.

But, to our knowledge, there is not yet a non-profit organization to support the *Preservation of Data Rescue Capabilities* for older digital assets (that vapor trail). Such a group would be of support to scientists, historians, librarians, and archivists. It would have its own stash, and collect details of other capabilities and projects across government, academia and commercial entities, and stay in touch, ensuring that the tools don't just fade away when projects end or certain staffers retire.

There are several other plans and goals for that group. If you'd like to learn more about that or send comments, we'd like to hear from you! Contact us at chris@mullermedia.com.

[Biographical Information]

Chris Muller has had the pleasure of working with researchers and archivists for many years, performing the "rescue" and conversion of digital legacy data. Long, happy dealings with NARA, IPUMS, state archives and several universities have made his work very enjoyable. His professional strengths include creating and enhancing recovery/conversion software, puzzling out arcane file formats and dealing with now-unusual media such as 9-track tape reels, as well as writing proposals and managing project teams. He is also proud to be a volunteer member of the CODATA Data-at-Risk Task Group, the InterPARES Trust PaaST Team, and the RDA Data Rescue interest group.