

PRESERVING ELECTRONIC ARCHIVES THE HARD WAY -  
THE AMERICAN CONTINENTAL CORPORATION RECORDS DATA RECOVERY

Robert P. Spindler  
Arizona State University Libraries  
1996 rev.

In 1994 Richard Cox postulated that a "paradigm shift" seemed to be occurring in the way archivists viewed the challenges and opportunities posed by electronic records. Indeed at least two major shifts in emphasis are evident in the professional literature: An expressed desire to establish strategies for preserving software-dependent computer files and creation of large scale research projects to develop functional record keeping requirements that can be used in the design of future information systems.<sup>1</sup> Although Cox, David Bearman and others have made important strides in developing functional record keeping requirements, there are few examples of successful models for long-term preservation of existing software-dependent electronic files.

This is the story of what happened when a traditional archival repository received its' first major accession of electronic records. I do not present this story as a model for managing software-dependent information, but I do believe it is a good example of the challenges we will all face as individuals and businesses donate their electronic legacies to the archives.

On April 14, 1989 as Lincoln Savings and Loan was being seized by federal regulators, representatives of the United States Bankruptcy Court and the Resolution Trust Corporation (RTC) arrived at the American Continental Corporation's office complex in Phoenix, Arizona and sealed off the perimeter. Government agents then proceeded to seize records of the corporation and its' subsidiary Lincoln Savings and Loan from every room of the complex. Records were placed in standard record boxes and identified with alphanumeric codes representing particular ACC departments. Records later recovered from the corporate warehouse were assigned a code for the warehouse, but not for their office of origin.

Soon afterwards Judge Richard M. Bilby of the US District Court, Arizona District at Tucson presided over the securities fraud and racketeering litigation collectively known as MDL-834. In the summer of 1989 Judge Bilby ordered the establishment of a document depository for the corporate records and other materials germane to the litigation in order to facilitate discovery by the numerous interested attorneys. The depository would eventually contain over 6,000 records center boxes.

As each box arrived at the depository lists of box numbers received and folder-level inventories of their contents were entered into a word processing file. The boxes arrived in random order and were inventoried at the folder level without precise references to the creators of the files. Box descriptions occasionally included identifications like "Mary M.'s files", or "Credenza in Bill's Office".

---

<sup>1</sup> Richard P. Cox, comp., University of Pittsburgh Recordkeeping Functional Requirements Project: Reports and Working Papers, School of Library and Information Science, University of Pittsburgh, Pittsburgh, 1994. Anne Gilliland-Swetland, "From Education to Application and Back: Archival Literature and an Electronic Records Curriculum", American Archivist, 56(3):532-545. David Bearman, The Implications of *Armstrong v. Executive Office of the President* for the Archival Management of Electronic Records, American Archivist, 56(4):674-689. Thomas R. Oglesby and William H. Leary, Managing Electronic Records, National Archives and Records Administration, Office of Records Administration, Washington D.C., 1990, pp.2, 8.

In 1990 the court ordered ACC to purchase a commercial document scanning and text indexing system to improve access to the corporate records. Hardcopy of the folder-level inventories for the ACC records was scanned into this system using optical character recognition technology, and documents selected by interested parties were scanned and maintained on an local area network based imaging system using optical disks as the primary storage medium. The imaging and indexing system was the first of a number of such systems designed by a commercial information management firm located in Phoenix.

As attorneys for both sides and officers of the court examined the materials each box was identified as containing information relevant to MDL-834, or irrelevant to MDL-834. Reviewers also had the opportunity to identify selected documents for application of attorney/client privilege restrictions. The scanned versions of the inventories for each box were then loaded into "Relevant", "Irrelevant" or "Privileged" data files and the hardcopy versions of those inventories were then bound into notebooks with the same identifications.

In June of 1993 the Department of Archives and Manuscripts at the Arizona State University Libraries received an inquiry from representatives of the American Continental Corporation (ACC) regarding a possible donation of the corporate records to the department's Arizona Collection. Ron Clifton, a consultant to the management team assigned to run ACC while it was under court supervision, invited department head Edward Oetting to the document depository to examine the collection. Oetting was shown the 6,000 boxes of material, the hardcopy folder-level guides, the LAN hardware that supported the imaging/indexing system and over one hundred optical disks containing digitized ACC records. The imaging system was not running at the time it was shown to Oetting, and the ACC staff was unable to provide any system documentation. Nevertheless, the hardcopy guides did provide accurate folder-level access to about 60% of the records. The paper guides occupied two record center boxes.

In July of 1993 Judge Bilby authorized ACC to continue negotiations with ASU to arrange for the transfer of the corporate records to the university at the conclusion of the litigation. Negotiations between the university and ACC attorneys continued through the summer of 1993, and in September the ACC LAN was loaned to the department in the hope that the imaging system could be repaired.

Richard Pearce-Moses, our Curator of Photographs who had acquired some significant knowledge of DOS and UNIX based systems, led the early stages of the departmental effort to recover the databases. He attempted to recover the system in its' original form, a local-area network (LAN) using a very early release of Novell Netware. When Pearce-Moses attempted to contact Novell for technical help over the phone, he was told that Novell did not support non-current releases, and that any further discussion would require a credit card number. During the winter and spring of 1994 he invited technical staff from the system designers and from the university Information Technology department to attempt to recover the database. ACC provided funding to hire a Novell certified LAN technician to examine the system, but upon his arrival he admitted he was unfamiliar with this release of Netware.

Pearce-Moses found himself caught in the crossfire between some experts who suggested this was a software problem and other experts who indicated the problem was with hardware. One hardware deficiency that was identified was that the Smile/AT target drive board was defective. A representative from Sanyo/Icon was nice enough to loan us a replacement board for the cost of shipping. Nevertheless, all of these recovery attempts failed.

On June 29, 1994 Judge Bilby delivered his court order approving the plan for selection and preservation of archival records and donation of the imaging system that was negotiated between ACC representatives and ASU. Judge Bilby's court order stipulated that portions of the collection that were described in the hardcopy guides were to be opened to the public by January 2, 1995. In August Pearce-Moses left the university in favor of another position and I was assigned responsibility for gaining physical control

over the collection and providing access to the collection through the hardcopy guides, and if possible through the imaging system, by January 2nd.

Since we had exhausted the technical support available to us, I followed up on some research Pearce-Moses started regarding commercial data recovery services. Ontrack Data Recovery of Eden Prairie, Minnesota was willing to examine the gigabyte external hard drive that came with the ACC system and make a data recovery cost estimate for about \$250. We packed up the drive and sent it to them for analysis. The data recovery firm examined the drive and indicated that although it had suffered some physical and electromagnetic damage the data could be salvaged. They provided a recovery cost estimate at about \$1,200. We asked the firm to send us a directory from the hard drive before we committed to the full cost of the recovery. On August 10th we received the first directory from the drive, which contained 415 megabytes of data in 1780 data files.

Closer examination of the directory and further discussions with the system designers revealed that the hard drive included software files, index data files and compressed image files in a number of formats. I faxed a copy of the directory to the system designers and called them with questions, but the system designers were unable to identify their own subdirectories and file names. They did remember that their early system prototypes used a commercial software package for the indexing portion of the system, and that the software linking the indexing to the digitized documents was their local code. They identified the subdirectories containing the indexing software called FolioViews, release 2.0, a text searching and retrieval program produced by Folio Corporation in Provo, Utah. However, the system designers indicated they felt it was doubtful that the linkages between the indexed text and the document images could be recovered. As I seemed to be wearing out my welcome with the system designers and they seemed to be unable to provide much additional assistance, I decided to concentrate on attempting to recover the text index portion of the system.

I immediately contacted Folio Corporation and they indicated that they were currently selling FolioViews 3.1 for Windows, and that data from release 2.1 may not be compatible with their current software. They told me that the software documentation was online, however they were able to find a quick reference manual in hardcopy for FolioViews 2.0, and they sent it to me at no charge.

My initial plan was to recover the scanned versions of the hardcopy inventories and set up that system for patron searching as a supplement to the two boxes of hardcopy finding aids. On August 10th I wrote in my recommendation to department head Oetting and to the Dean of Libraries Sherrie Schmidt "Failure is still a possibility since we are working on speculation by the system designers about the identity of these files, since long term support for FolioViews 2.0 is doubtful and upgrading to newer software could corrupt the data."<sup>2</sup> To her credit the Dean supported me and authorized expenditures for the data recovery.

At the end of August the data arrived on seven 60 megabyte tape cartridges. Using the tape backup facility in my 386 PC I was able to identify and copy the subdirectories containing the FolioViews software and a couple of sample "infobases", the FolioViews name for the indexed text files. I loaded the software and attempted to bring up one of the test infobases, and I was greeted with the admonition "Error - Incompatible Infobase". Soon I was back on the phone with FolioViews, and they suggested that the infobases may have been produced in an earlier FolioViews release. I then discovered FolioViews online documentation and found that the software included an upgrading utility that converted FolioViews version 1.3 to FolioViews 2.0. I made a copy of the file, crossed my fingers and invoked the upgrade utility. The five megabyte file was successfully converted in about forty minutes.

I opened the infobase and discovered that the software manipulated the upgraded datafiles correctly. I

---

<sup>2</sup> Spindler to Edward Oetting and Sherrie Schmidt, August 10, 1995. Spindler administrative files.

tested a number of the software search and display functions as I learned them by reading the software documentation. It became evident that the text files contained a significant number of errors, which I attributed to the use of the OCR technology in scanning the paper inventories. I then spent a significant amount of time comparing the text file to the hardcopy inventories to assess the completeness of the electronic version. Since many of the box numbers for the collection contained OCR scanning errors, and since the database could not serve as an accurate finding aid for the collection without accurate box numbers, I personally examined 1722 pages of electronic Relevant inventories and corrected the electronic box numbers by comparing the electronic and hardcopy inventory texts. 23 of the 1722 relevant inventory pages were corrupted beyond recognition. In the end, 87% of the Relevant series hardcopy inventories were available in the Relevant infobase.

The situation with the Irrelevant series records was very different, since we had previously appraised and selected 193 of the original 1200 Irrelevant boxes for the archival collection. I photocopied the pages of the hardcopy inventory that coincided with the 193 boxes we had selected and then compared the hardcopy against the irrelevant infobase. I was able to locate 100% of the infobase pages for the archival Irrelevant files and copied them to a new infobase that would serve as a finding aid. For both the Relevant and Irrelevant series, I did not attempt to correct all the OCR errors at this time, concentrating on verification of the box numbers and creation of infobases that matched the hardcopy guides as much as possible.

By December we had a complete hardcopy finding aid for Relevant and Irrelevant files and a database that allows truncated keyword searching with proximity control for the vast majority of the collection. However, cleanup of the full text of the inventory pages is still in progress to ensure consistent and comprehensive text searching. Appropriate warnings about the OCR errors and the scope of the infobases are included in the hardcopy finding aid for the collection. I gave a regrettably brief orientation/training session for the reference and retrieval staff, and transferred the files to the reading room patron access computer. The collection was opened to the public on January 2, 1995 in accordance with Judge Bilby's court order.

Nevertheless, the department still faces a number of challenges with respect to its responsibility to appraise and preserve the electronic archival information. Of the 1780 data files recovered from the external harddrive, only 46 have been appraised, and most of these are FolioViews program files. So far we have been unable to marshal the resources to read the one hundred optical disks, even though we have the two original optical disk drives. We do not currently have sufficient disk space to download the optical disk information if it is in fact archival. At this writing the FolioViews 2.0 software is running happily on three of our departmental machines, but continued support for 2.0 from the manufacturer is unlikely, and upgrading may corrupt the data. In addition, the tape backup system we used to extract the data from the tapes we received from Ontrack failed. When we went to our campus supplier to replace the tape drive, they had discontinued that model and had no compatible models available. Our Library Technology staff fortunately has a working tape drive that is compatible with the ACC tape cartridges, and we will soon be copying the data over to tapes compatible with our new drive using their facilities.

In retrospect we faced a number of external and internal forces that will undoubtedly be faced by other archives that acquire electronic records. The external issues include inconsistency between software manufacturers in their willingness or ability to provide support for old releases of their products. Specialized vendor-specific training of technicians does not enable them to effectively address problems in systems employing a variety of software and hardware products, resulting in the "It's not the hardware, it's the software/it's not the software it's the hardware" syndrome. Absence of system metadata or total reliance upon online system documentation is a serious impediment to our ability to preserve electronic information since the system may not be running by the time it is sent to the archives.

The internal issues are numerous and substantial. Acquiring the technical skill to deal with these materials continues to be the principle challenge facing our repository. However, there are a series of basic skills that I have acquired that enabled us to understand and assess our electronic holdings. By establishing

backup routines, data dictionaries and policies and procedures for handling our electronic administrative information, I learned many of the procedures needed to manage the ACC files. As a result we've been able to acquire and maintain the ACC information, and make some determinations as to where we'll need external assistance and support. In order to begin managing electronic information produced by others, archivists can learn a great deal from their experiences managing their own electronic information. Nevertheless it is important that archivists and their supervisors recognize the limits of their knowledge, and the potential damage that pushing those limits can cause.

However, archivists cannot be expected to keep abreast of all the information technologies that are likely to be used by potential donors. David Bearman wrote "Recordkeeping is not the province of archivists, records managers, or systems administrators alone but is an essential role of all employees and of individuals in their private lives."<sup>3</sup> Ultimately, archival appraisal of existing electronic records systems may best be conducted by information analysis teams composed of archivists, historians, computing professionals and users of the original systems.

An important issue in the ACC project was trying to determine which was the most complete version of a given file. There were a number of infobases with duplicate file names in different subdirectories, some of which were probably backup copies. I chose to select the infobases from one particular subdirectory since they had the largest file sizes of the duplicate versions. This was not a reliable or scientific selection method. A related issue is that I have been unable to determine if the ACC files have been intentionally damaged or corrupted. Archivists need to develop strategies for authenticating electronic information, and mechanisms for preventing unauthorized tampering with the archival information once it is authenticated.

Maintenance and upkeep of our computing equipment and infrastructure has been the departmental responsibility no one wants but everyone needs. Even though we have fifteen microcomputers in three physically separate office areas we have chosen not to purchase a local area network because of the costs and complexities of its care and feeding among other factors. We use File Transfer Protocol (FTP) or floppy disks for our file sharing and transmission. This is indeed a low technology alternative but it is technology that we can for the most part support at this time.

Unfortunately this is also changing. Our ability to secure the resources we need to preserve electronic information is lagging far behind the amount and complexity of the archival electronic information that has already been created, and will soon be left on our doorstep. Archivists cannot preserve electronic information on systems that are inferior to the systems that originally created the information. The gap between the technology that is available in home or office PC environments and the technology used by archives and libraries is widening, and this trend is likely to continue as long as libraries and archives are unable to acquire technologies common to many homes and offices. As a result the most effective models for the future may be cooperative arrangements for data storage and maintenance with information professionals from other departments or institutions that can justify acquisition of new technologies. We need to start developing those relationships now.

Finally, we are concerned about the proliferation of technologies our department is already facing. Our reference area now contains databases in three software programs, two CD-ROM indexes, a terminal for access to the University Libraries LAN, and our online catalog. This constitutes serious challenges in terms of staff development and training and user instruction. This proliferation is likely to continue since archivists now believe that in many cases it is the combination of the data file and the software that constitutes the record.<sup>4</sup> As

<sup>3</sup> Bearman, "Implications of Armstrong...", pp.685.

<sup>4</sup> See David Bearman "Archival Methods", *Archives and Museum Informatics Technical Report*, 3(1): 28. (1989) and Jeff Rothenberg, "Ensuring the Longevity of Digital Documents", *Scientific American*, 272(1):46-47. (1995)

a result archives are beginning to face a proliferation of platforms and environments used by their donors and their staff. In the short term reference staff must be prepared to interpret all these systems for users.

In the long term the problem of environment proliferation could diminish as various forms of object-oriented programming and operating system-independent systems emerge. Jeff Rothenberg's recent *Scientific American* article points out the potential for creating systems with built-in software emulations that mimic the functions of old software. But until Rothenberg's dreamware comes to life we have 15-20 years of records created in software-dependent environments that could be lost.<sup>5</sup>

Our experience with the American Continental Corporation Records has forced us to address a number of issues relating to appraisal and management of archival electronic information. However, this experience has been most valuable as a demonstration that we have only touched the tip of a very large virtual iceberg.

---

<sup>5</sup> Jeff Rothenberg, "Ensuring the Longevity...", pp.47.