Save File
By Charlotte Abney Salomon, Ph.D student, History of Science & Medicine

Yale's digital preservationists work to keep aging media readable
The writer Tony Geiss helped to create some of the most iconic children's entertainment of the past thirty years, co-writing "An American Tail," "The Land Before Time," and thirty-six years of Sesame Street. During his career, he invented the Muppet monsters Abby Cadabby and the Honkers, co-created the segment Elmo's World, and shared 22 Emmy awards for scriptwriting. Geiss died in 2011 at the age of 86, and in 2013, the Beinecke Rare Book and Manuscript Library at Yale bought a collection of his creative materials, including notebooks, recordings, scripts, and stuffed animals.[i]

The executors of his estate also donated Geiss's Apple Hard Disk 20SC, a twenty-five-year-old external hard drive about the size and shape of a cereal box containing several years' worth of scripts, projects, and personal files from the late '80s and early '90s. While the library staff had no trouble cataloging the papers and making them available to researchers, it quickly became apparent that the work on the hard drive might be irretrievable due to its age. The task of finding out what was on the drive and preserving it fell to Gabriela Redwine, Beinecke's digital archivist.[ii]

Redwine knew that just turning on the drive might alter its contents unless she used special software, but the only computer at the library with the right connector to use with the drive was a mid-'90s Power Macintosh, too old to run the protective program. She carefully made copies of the files she could access, but what she really needed to do was to "image" the drive, to make a full copy of the entire disk. The drive wasn't spinning all the way up, though, meaning not all of it was accessible, and Redwine worried that some of its contents might never be retrieved. "When a disk doesn't spin up completely, it's kind of a mark of it degrading and failing," she says.

Using the disk utility software on the Power Macintosh, she tried several times to image the drive. "The imaging process would start," says Redwine, "but it would just poop out in the middle because the drive wasn't fully present to be captured, essentially. And then, one magical time, it was." Seizing her lucky moment, she saved a copy of the disk, knowing it may have been her only chance. "It's probably too damaged to ever be able to get a disk image again," she says.

The disk contained hundreds of files, including dozens of scripts for Sesame Street. One file, named "2085," the number of the 1991 episode for which it was written, holds a script titled "FABULOUS RAINBOW-TG." Its first lines are for Big Bird, who says to the camera, "Hi! Welcome to Sesame Street. We just discovered something. We're each a different color." Some of the scripts differ slightly from the televised episodes, and this hard drive may hold the only existing copies of these early versions.

Euan Cochrane, Digital Preservation Manager for the Yale Library system, was working with Redwine the day the drive imaging succeeded. "I think we'll probably never be able to interact with that drive again," he says. "I think we got the last chance to get the data from it."

Tony Geiss's hard drive is one of thousands of pieces of digital media found in the libraries at Yale, including documents, programs, video, music, and games created and stored as computer files. With each new version of software programs and each new generation of computer hardware, older files and models are left behind, incompatible with anything on the market. While the library's books may fall apart after a century or two, computer files are often unreadable after less than a decade.[iii] Over the past few decades, the new discipline of digital preservation has formed to combat this trend. Cochrane,

Redwine, and specialists like them are working to save digital files, often one at a time, from being lost to obsolescence.

For as long as people have stored information by writing on physical objects, text has only remained readable so long as both the physical object and the language of encoding have survived. Viking-age runestones are still readable thanks to the durability of the stone and the continuous human knowledge of Old Norse.[iv] Ancient Egyptian hieroglyphs, though physically unchanged, languished in unreadability until the Rosetta Stone was translated in the early 1800s.[v] Over the last millennium, most writing has taken the form of fragile paper books, but with careful storage and modern conservation techniques, many have been preserved for centuries.

The advent of personal computing in the late 1970s seemed to promise an end to worries about preservation. Computer files are endlessly reproducible. Unfortunately, digital writing has turned out to be much less permanent than the paper it has largely replaced; nearly everyone today has a disk of files or a video game cartridge that is simply unreadable, either because it is damaged or because no hardware will open it anymore.

In January of 1995, writer Jeff Rothenberg published an article in Scientific American magazine alerting readers to the pressing need for digital preservation.[vi] In it, he imagines his hypothetical grandchildren in the year 2045 unable to open or decipher digital files stored on a 1995 CD-ROM, while an accompanying letter in an envelope is immediately readable. “Unfortunately, many of the traditional methods developed for archiving printed matter are not applicable to electronic files,” Rothenberg wrote. “The content and historical value of thousands of records, databases and personal documents may be irretrievably lost to future generations if we do not take steps to preserve them now.”

Cochrane credits Rothenberg’s article with founding the discipline of digital preservation, but even now, he says, the field is still grappling with the basics. “That's not even twenty years ago,” he says. “People are still just trying to figure out what we need to be doing. There's still not a consensus about what the best way to approach all of this is.”

Within the Yale library system, digital materials are now getting targeted attention. The 15 libraries at Yale together house around 15 million items in all formats.[vii] The staff of the library’s department of preservation, established in 1971, draws on decades of science and experience to preserve paper books of all ages, but saving digital media required a specialist.[viii] In 2013, the library hired Cochrane in the newly created position of Digital Preservation Manager. A former digital preservationist for the national archives of his native New Zealand, Cochrane now manages the rapid aging of the "born digital" materials the library is steadily acquiring.[ix]

"What I'm charged with doing is setting up all of the services we need to be able to preserve digital content for as long as it's needed," says Cochrane.[x] Within a normal cycle of technological updates, he says, files and hardware only stay current for between five and eight years. Once a file or disk has been around longer than that, it normally becomes unreadable on any current computer, and preservationists have to come up with an alternative way to open it.

The Yale libraries hold digital materials in a variety of formats, stored on a variety of media dating back decades, including Zip disks, CD-ROMs, three different sizes of floppy disks, many kinds of magnetic tape. They have also acquired multiple hard drives like the one from the estate of Tony Geiss. "Those can have anything on them,” says Cochrane. “They can have office documents. They can have image

files. They can have audio. They can have databases, which are complex and challenging to get hold of. Any type of digital file you can think of."

Even at Beinecke, which usually acquires the papers of older writers at the end of distinguished careers, collections are including increasing numbers of digital files. "We haven't yet gotten anything from anyone who is what the media call a 'digital native,' says Redwine. "But we're getting more and more digital every year, so we're definitely moving in that direction."

Cochrane's first priority is to find, identify and store the digital materials filed away in the library's collections. Many of them were never properly cataloged, and Cochrane has no idea how many uncounted disks are hiding between books on various library shelves. "What we probably ought to do in the long term is go through and have someone physically look everywhere," he says.

Cochrane's office sits within a tile-walled warren of underground workrooms and hallways spread below Yale's Sterling Memorial and Beinecke libraries. His desk features futuristic computer monitors of different sizes, and in front of it stands a library cart packed with plastic cases holding CD-ROMs and 3.5-inch floppy disks. He has pulled this collection randomly from the library stacks several floors above to get a sense of what kinds of digital files he is working with. "There's some weird stuff in there," he says. "Half of it's commercial, but half of it is that some professor has written something to a disk or CD-ROM and it's probably the only copy anywhere." He estimates that there are between six and eight thousand CDs and floppy disks in the Sterling library stacks.

Each disk may hold any number of files of different sizes and types. "Out of the first two hundred that we took an image of," Cochrane says," there are about 173,000 files, just on those few CDs, in 136 different file formats with 43 different versions of those file formats." If this is a representative sample, Sterling's collection totals around four terabytes of data, but this is only one library. "We think at the moment," he says, "that the library system - Beinecke, the main library, and the other ones that are part of the main system - have about a petabyte of data, which is a thousand terabytes. It's a million gigabytes, roughly." Equivalent to over 13 years of high-definition video, a petabyte is a massive amount of data.[xi]

Since the physical objects on which digital files are stored degrade easily, Cochrane, Redwine, and other digital archivists concentrate on moving data safely into storage so that it can be retrieved later without changing it. The library currently stores one copy of each file it retrieves on servers and data tape drives on campus, and in the next few years it will begin storing a backup copy of each file somewhere outside of Connecticut in case of a natural disaster. In each collection, preservationists must continually organize, manage, and track the files.[xii]

Cochrane says the biggest challenge in large-scale storage is finding consistent funding, because the amount of data to save is always growing. "The difference between building a physical building and a digital space is that you may not invest as much up front, but every year after that you're going to invest more," he says. "It's easily in the millions every year." Despite the expense, Cochrane says that delaying building the data storage would be risky for the library. "There's a good chance that we'll lose stuff that we have if we don't have somewhere to put it that's secure. We'll also lose the opportunity to bring new collections in."

Moving a file into storage starts with finding a compatible computer and drive to read the disk. Many are still readable on library computers, but older ones require vintage hardware, like the Power

Macintosh that Redwine found to read Tony Geiss's hard drive. For another recent project, says Cochrane, "we had to buy some floppy drives off of eBay, because you can't buy them anymore."

To make this step easier, Cochrane is building a collection of vintage hardware of different ages, stored in an equipment cage in a basement storage room down the hall from his office. Beige monitors and a candy-blue iMac CPU sit on metal shelves beside cardboard boxes filled with portable drives and unopened packages of blank ZIP disks and 3.5-inch floppies.[xiii] "We're planning on using it for all sorts of stuff, including testing different preservation approaches against the original in the original context," says Cochrane, "but also in some cases we may have to have the hardware just to be able to get access to it."

Even with the right hardware, it still takes compatible software to open individual files, and this can be an even bigger challenge. "If you're really lucky you might have some software that can interpret them, and you can open them," says Cochrane. "That's the best outcome." Sometimes, though, they aren't so lucky. For example, the library has an archive of AutoCAD design files from the architecture school that don't open in current versions of the software. "Someone had contacted AutoCAD about getting their old software, and they said they don't even have it anymore," says Cochrane.

Once library archivists have safely opened and stored the files, their next goal is to make the files available to patrons and researchers using their own personal computers and current software. Those of us who only manage our own files often assume that the solution is simply to convert each old document by opening it in a current program and saving it in the new format, a process called migration. "People get that," Cochrane says, "because they've done it themselves, and it all just seems pretty simple." It's not.

"The trouble with migration," says Cochrane, "is that it can alter the content, or you can lose content." When Redwine first opened the Sesame Street scripts on Tony Geiss's hard drive, for example, the computer she used converted them to a different file format, which changed the spacing of the text. She says this is a problem she deals with frequently. "Look and feel may not matter for the majority of the documents you're working with," she says, "but if you're working with the work of a digital poet, then a line break does matter a lot. It can be the difference between drafts or a final version or an editorial comment."

The only way to open old files in their original format is to use the software from the time in which it was created. Both the software and hardware that can run it are difficult for librarians to find and store and would be wildly impractical for library patrons to check out in order to study the files. Instead, preservationists have turned to software programs, called emulators, that mimic older hardware, presenting the user with a simulated vintage computer that can then open vintage files within vintage programs.

The first emulators were developed in the late 1980s and have been used ever since, largely to play console video games on personal computers.[xiv] Digital preservationists are now using emulators to essentially recreate entire computers from specific times in order to open files in their original formats, without the changes that come from migration.

Earlier this year, Redwine and Cochrane set up an emulator called Mini vMac to open the disk image of Tony Geiss's Apple Hard Disk 20SC. In order to use it, a reader checks out an up-to-date Dell laptop at the library counter. Among seven files on the desktop labeled with the names of writers, one is labeled

“Geiss, Tony.”

"To run the emulator, you just double-click on Mini vMac," says Redwine, clicking the third icon inside the file. "You can't screw this up, by the way." On the laptop screen, a four-inch-wide gray window appears, and in the center is a pixelated image of a classic Mac with a blinking question mark. She opens the file containing the disk image for the hard drive, and within the window appears the solid black home screen of Geiss's personal computer.

Its folders are all clickable, and each file opens within the word processor it was created in, fully formatted. In addition to the dozens of Sesame Street scripts, there are other documents Geiss created, including a letter volunteering his creative services for the presidential campaign of Bill Clinton in 1992 and a letter to the editor of the New York Times critical of President George H.W. Bush. One small file called “bio tg” contains a brief, one-paragraph autobiography that describes his work in adult television, for comedians like Dick Cavett and Robert Klein, alongside his writing for kids. It concludes, “He has won seven Emmies as a Sesame Street Writer, and a Prix Jeunesse as a writer of ‘Don’t Eat The Pictures: Sesame Street at the Metropolitan Museum.’ He is a muppet.”[xv]

Cochrane sees using an emulation of a writer’s virtual desktop, with all of its extra details about the writer’s life and work, as a more immersive look into the writer’s world for a researcher than reading paper notebooks. “It's a whole new experience that we were never able to give for our physical materials," he says.

Since the technology is new, there were surprises for the preservationists as they created the emulation. "One of the things that Euan and I discovered," says Redwine, "is that when we created this disk image and then put it in the emulator and started interacting with it, we realized that we were changing dates. The file date would change whenever we opened it." She has since locked the files so that researchers will not inadvertently change the hard drive image by viewing it. Other things can't yet be changed at all. "I haven't figured out a way to make the window bigger," she says, laughing. "You not only experience the look and feel, you experience the frustrations of the look and feel."

A dedicated emulation can be an ideal solution for an individual file, a collection of files, or a single hard drive, but each emulator can only provide one simulated computer in one configuration. The emulator for Geiss's hard drive, for instance, would not be able to open files created even a few years before or after, or created on a Windows PC. Setting up each emulator also requires intricate coding. "Emulation software has traditionally been seen as something that's complicated and difficult to handle," says Cochrane.

Under his management, however, Yale is helping to develop a brand new, user-friendly form of emulation, which runs vintage software through a web browser.
While working at Archives New Zealand several years ago, Cochrane and a visiting German computer scientist came up with an idea for making emulation easier to use by running it where the files are stored rather than installing it on each user’s computer. Cochrane brought the partnership and the project with him to Yale. “We now at the moment have the only installation of this outside of Germany,” says Cochrane. “Because I was there at the beginning when they were doing this, they were happy to work with us.”

The system is called Emulation as a Service and was created within a collaborative project called Baden-Württemburg Functional Long-Term Archiving and Access (bwFLA) at the University of Freiburg.[xvi] The

program combines a wide variety of different emulators that each run within a Web browser instead of being installed on a computer. "The idea is that we take on all that configuration stuff as the experts," Cochrane says, "and then the end users can just see it and interact with it, and they don't need to think about all the difficulty." Each time a user selects a file, the program provides the right emulator loaded with the right software and opens them all at once within the browser window.

"What we'd like to have is preconfigured entire environments that have, say, Windows, plus Office, plus CD add-ons or extensions that you'd need," says Cochrane. Eventually, he imagines adding a link to the emulator software next to each digital file in the online Yale library catalog. "You click on it and it opens the environment, with the file attached, and it opens within that software."

Cochrane opens up a browser window on his computer and clicks through to his copy of the online emulation. Within the window appears the desktop of a computer from the late 1990s. Inside it, he opens the list of applications and chooses the program *Mavis Beacon Teaches Typing*. Within seconds he is playing a touch typing game just as it would have appeared at the time it was created.[xvii] "It's an awful lot of work, and the emulation's not perfect yet," says Cochrane. "But it's early days."

There are several demos of the emulation service on the group's website. For the moment, however, the emulation service for Yale materials is not available to library users or anyone else outside of Cochrane's office. Before Cochrane can open it up, the the university must legally acquire the rights to run three decades' worth of operating systems and programs all at once on a virtual computer. "You'd need to have that catalog of software environments preconfigured and available to associate with those files when you bring them into the file server," says Cochrane, "and we don't have any of that right now."

It isn't even clear what kind of licenses would apply to an online emulation service. "If you go into an old Windows license," says Cochrane, "it says you can have it on one PC. It doesn't say what a PC is, and it doesn't say that you can have it on a virtual PC or anything like that."

Cochrane says that he's in talks with Microsoft and is thinking about contacting Apple next. But in some cases, even if the company is willing to license the software, it may not be able to provide it in the first place, as with the unreadable AutoCAD files from the architecture school. "We'd probably have to find some on eBay or something like that, and then get permission from AutoCAD to reuse the old software, and probably have to pay them something for it. But we're going to try." He hopes to begin to produce the program at Yale next year, but there is no official plan to make it public yet.[xviii]

For Cochrane, the key to a long-term solution for the problem of digital preservation is for users to create and edit files in open-source software rather than using programs that require licenses. Open-source software, as its name suggests, is freely available, as is the code in which it is written, which lets digital preservationists use and modify it at will, avoiding nearly all of the hurdles they currently encounter. "If we need to replicate what it was like to interact with that content at the time," says Cochrane, "we can just take that software and use it without having to worry about it, and if we wanted to we could potentially recompile it for new operating systems because we'd have access to the source code."

He acknowledges that proprietary software, like other trade secrets, may always be necessary for companies to be able to develop and profit from new technologies. Still, he hopes that some accommodation can be made. "I'm of the opinion that we should have solutions for all possible things

that could be happening, so even if people come up with these weird proprietary formats we should have a way to preserve them anyway," he says. "If it's helping you to build your business, great. We need to figure out what we're going to do with that rather than just saying you're wrong for having done that."

Still, for the sake of preserving our digital documents for future generations, Cochrane recommends making the switch to open source software. "People creating stuff using open standards is going to make it easier for us to access it in the future."

---

[i] All Geiss biography: http://www.nytimes.com/2011/01/31/arts/television/31geiss.html

[ii] http://connect.clir.org/blogs/gabriela-redwine/2013/10/31/born-digital-planning-for-access ; interview with Redwine at Beinecke, 11/26/14. All Redwine quotes from same interview.

[iii] http://www.clir.org/pubs/archives/ensuring.pdf

[iv] http://www.runor.se

[v] http://www.ancientegypt.co.uk/writing/rosetta.html

[vi] http://www.geol.queensu.ca/faculty/harrap/teaching/geol463/downloads/files/Rothenberg_Longevity_SciAmer1995original.pdf ; interview with Euan Cochrane, 9/17/14

[vii] http://guides.library.yale.edu/about

[viii] https://collaborate.library.yale.edu/lhr-public/jobs/digitalpres.aspx ; personal tour of preservation department, 6/25/14

[ix] https://collaborate.library.yale.edu/lhr-public/jobs/digitalpres.aspx

[x] All quotes from Cochrane: Interviews with Euan Cochrane, Sterling Memorial Library, 9/17/14 and 11/11/14

[xi] http://gizmodo.com/5309889/how-large-is-a-petabyte

[xii] New information this draft: Email from Cochrane, 12/10/14

[xiii] Interview with Cochrane, 11/11/14

[xiv] *The Emulation User's Guide* by Kenneth Stevens, page 58-64, found at: http://books.google.com/books?id=HQB2AgAAQBAJ&pg=PA57&lpg=PA57&dq=history+of+emulation&source=bl&ots=WxjsDNsurD&sig=j4-TOcTDt__byS3cBwxeFoue-xY&hl=en&sa=X&ei=JMyFVKOkIsGqNoyXgyA&ved=0CGEQ6AEwCQ#v=onepage&q=history%20of%20emulation&f=false

[xv] All descriptions of Geiss's hard drive and emulation: Personal use at Beinecke, 11/16/14

[xvi] http://bw-fla.uni-freiburg.de

[xvii] In this version, I have replaced the demonstration he gave me on 9/17 with the one he gave on 11/11.

[xviii] Email from Cochrane, 12/10/14