Addressing the Messiness of Electronic Records Acquisition: Discussion of Methods and Proposed Professional Directions

Cal Lee
School of Information and Library Science
University of North Carolina, Chapel Hill

Electronic Records Section
Society of American Archivists Annual Meeting
August 14, 2009
Austin, TX

UNC
SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Professionals have a special role within society.

Professional status affords unique privileges (e.g. lawyers give legal advice, doctors write prescriptions, pilots can step behind the cockpit door).

There are two fundamental justifications for these privileges:

(1) Their activities and decisions are based on a distinct body of expertise.

(2) They have agreed to use their professional status to act in the public interest.

Historically, individuals and families have accumulated and managed personal archives.

Most of these collections have been relatively small and haven’t left the homes of the collectors.

Collections of a few prominent individuals and families made the transition into collecting institutions.

Many cultural institutions were initially seeded by personal collections of influential people.

Over the past half century, **four trends** have radically changed the nature and status of personal collections.

First, work within collecting institutions has become increasingly **professionalized**.

Specialization

Professional education and training (available and expected)

Conferences

Journals

Professionals associations

Specialized language and secret handshakes

Second, individuals have gained more **ability to create and store** materials that they find meaningful, useful, or simply more convenient to keep than to discard.

Third, **researchers** have placed considerably more emphasis on the importance of personal stories, voices and perspectives.

Finally, previously distinct communities have come to recognize that they **share challenges**, associated with long-term care of digital resources...

Digital objects are created and perpetuated through physical things (e.g. charged magnetic particles, pulses of light, holes in disks), but...

They are **not** quite like spatio-temporal objects (regular, physical things).

Digital objects are sets of **instructions** for future interaction

Interactions require numerous technological components to **come together** at the right place and time.

What are the **implications** of these trends for the place of archivists in the realm of personal digital archives?

Remember: We trust professionals to do special things, because...

(1) They have (or should have) special **expertise**

(2) They promise to act in the **public interest**

So how about that distinct body of expertise?

What’s required to “do” digital curation?

Reflecting **purposes** – understanding and attending to intentions of creators and “primary users”

Avoiding Unnecessary Lock-In

- “How do I get this stuff out when I stop using this particular system?”
- Hint: “No worries. You’ll always be using this system.” is **not** the right answer

Unlocking from Lock-in

- Getting files off removable media without doing irreversible harm
- Dealing with obsolete formats and platforms

Promoting **Discovery** through Intellectual Control

- Description
- Naming conventions
- Mappings across inconsistent terminologies

Promoting **Sensemaking**

- “Now I have it, but what am I looking at?”
- Creating, capturing, or extracting information for making sense of things being used

Acting Locally, but Thinking Globally

- To whom might I hand these things off in the future?
- How would that work?
- What are the likely motivations and needs of the recipient?

Ok, those sound like good things.

But what will it mean to act in the public interest?

Make sure that these things **get done** in socially responsible ways

This is not the same as **doing** all of these things

Don’t assume – *a priori* – who will be doing particular things

Strive to continually push the **frontier** of what's possible, but...

Honestly disclose what we believe we're really able to promise each other.

Practice “respectful and informed ignorance”

Will Rogers said, "Everybody is ignorant, only on different subjects.”

We need to bring our own answers and **informed** questions to the conversation

So what specifically are the roles we can play as professionals?

Get, Grab, and Guide

Get
(On Removable Media)

Many archives have received entire computers from record creators & donors.

Even more common is the “disk in a box” – floppy disks, CDs and other removable media among the physical materials obtained along with primarily analog collections.

The media will inevitably become unreadable (if they haven’t already).

Archivists must extract whatever useful information resides on the media, while avoiding the accidental alteration of data or metadata.

The field of digital forensics is a source of expertise, principles, methods and tools for archivists, including...

1. Recovering data when layers of technology fail or are no longer available:

- Reading, analyzing and manipulating hex dumps of files
- Recovering data from temporary, unallocated and slack space
- Identifying file types through automated analysis of file content (e.g. headers & file signatures)
- Guessing passwords & breaking encryption

2. Capturing evidence from places on a computer system that are not always immediately visible, e.g.:

- User account information
- Files on disk used for virtual memory management
- Temp files
- Various caches
- “Recent documents” in Windows
- Cookies
- History files
- Configuration files (often from Registry in Windows)

3. Ensuring that actions taken on files don't **unintentionally** make irreversible changes to essential characteristics

Examples of Irreversible Changes

- Lossy compression (e.g. JPEG)
- Lower-quality surrogate (e.g. thumbnail image, access copy of video)
- Format conversion (e.g. Word to PDF/A, Excel to CSV)
- Character encoding (e.g. EBCDIC to ASCII)
- Normalization of data values (e.g. date values in a database to a common date encoding)
- Rewriting pointers (e.g. links in a web site from absolute to relative or vice versa)
- Overwriting older versions files or values with newer versions
- Pulling files out of their native file system

Strategies for Avoiding Accidental Manipulation of Volatile Data

- Use write-blocking equipment when first reading from a medium (hardware, if possible)
- Make bit-level images of storage media
- Create checksums before and after file transfers and transformations

4. Attending to Order of Volatility

- Some types of data change much more quickly & often than others
- Important to recognize in order to recover data from a computer system or media, while ensuring that actions don't make irreversible changes to their record characteristics
- Example: If the contents of the browser cache are important to you, capture the cache before using the browser

5. Taking advantage of a wide array of tools & techniques already available

- Digital forensics literature, training & events
- Free & open-source tools (e.g. AFF, The Sleuth Kit)
- Commercial packages (e.g. EnCase, Forensic Tool Kit)

6. Adopting established practices for documenting what we do, so others will know what we might have changed

Guidelines for Evidence Collection & Archiving (RFC 3227) – Some Highlights

- “Keep detailed notes.”
- “Minimise changes to the data as you are collecting it.”
- “Do collection first and analysis later.”
- “Proceed from the volatile to the less volatile.”
- Computer evidence should be: admissible, authentic, complete, reliable, believable

7. Recognizing & confronting the **ethical** implications of obtaining & providing access to data that reside at various levels of representation

Grab (From the Internet)

With the adoption of highly interactive Web technologies (frequently labeled “Web 2.0”), forms of individual documentation and expression are often also inherently **social** and **public**.

Such online environments do allow for personal documentation that is comparable in many ways to previous forms of personal documentation.

But...

They also engage external audiences in ways not previously possible.

Many individuals now rely significantly on “cloud” services for creating, managing & sharing records.

This poses numerous risks.

Sustainability Risk factors of Reliance on web Service Providers

- Expiration of service
- After a period of user inactivity, data deletion is triggered
- Changes in service offerings
- Companies going out of business
- Take-down based on complaints from other parties
- Mergers resulting in major displacement or complete loss
- Accidental loss due to drive failure & insufficient backup
- Purposeful destruction of data by malicious attackers

So “grabbing” at-risk records from the Web is one important approach.

Completely seamless access and management of all one's personal information is neither probable nor desirable.

So get used to the idea of *fonds* that are impartial, inconsistent, unintegrated, and otherwise downright **messy**, but also...

inundated with traces of individual lives, can be rich sources of meaning and are likely to be at risk.

Parts of someone's *fonds* can build on each other.

For example, if given some a person's digital "papers," one could find pointers to more of her online presence from the Web, her email or the storage media used by her computer, especially the hard drive.

See:

Culotta, Aron, Ron Bekkerman, and Andrew Mccallum. "Extracting Social Networks and Contact Information from Email and the Web." Paper presented at the First Conference on Email and Anti-Spam, Mountain View, CA, July 30-31, 2004.

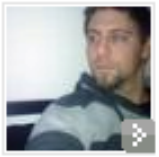
Garfinkel, Simson, and David Cox. "Finding and Archiving the Internet Footprint." Paper presented at the First Digital Lives Research Conference: Personal Digital Archives for the 21st Century, London, UK, February 9-11, 2009.

Guide

Archivists will only ever have custody of a tiny sliver of the documentary traces of individuals.

So helping them to curate their own materials can be just as important as taking custody of them.

So maybe we should get in on things like this:

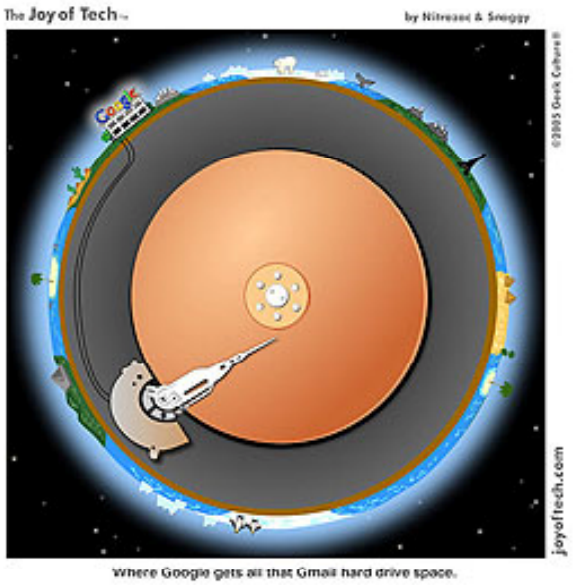


HOW TO: Take Your Data Back From Google's Claws

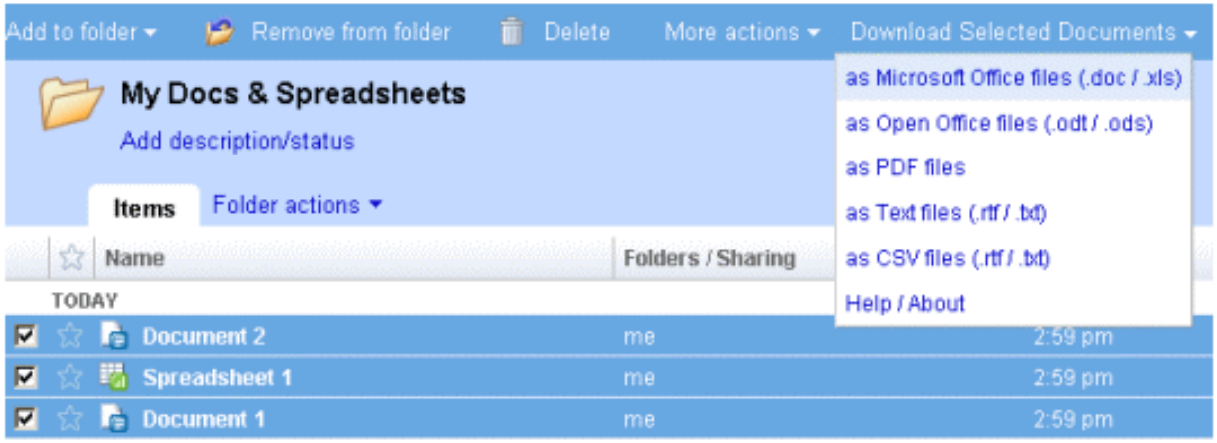
February 2nd, 2009 | by **Stan Schroeder** — 45 Comments

We've all pretty much become accustomed to the notion that Google :D is this invincible internet giant which will always be there for us, but it's not always true. A good example was this weekend's fiasco, when (due to human error) Google's search engine reported all web sites on the internet as unsafe.

Let's face it: **every** web service, Google included, can mess up, and sometimes it means losing your data. So, when was the last time you backed up the data on the various Google services you use? I thought so. Let's look at some easy solutions for extracting and backing up your data on popular Google apps and services.



Google Docs



http://mashable.com/2009/02/02/google-backup/

And this:

TOSBack

The terms-of-service tracker.

A project of the EFF

Hello

Learn About TOSBack

Subscribe to RSS

Highlighted Policies

Blizzard World Of Warcraft Terms Of Use

EBay User Agreement

Facebook Privacy Policy

GoDaddy Universal Terms Of Service

Organizations

Amazon

Apple

Automattic

Blizzard

Craigslist

Data.gov

DoubleClick

EBay

TOSBack keeps an eye on 44 website policies.
Every time one of them changes, you'll see an update here.

Facebook
changed its Privacy Policy
June 2 around 5pm PT

TOSBack started tracking a new policy.
It's the Google Blogger Terms Of Service.
June 2 around 2pm PT

TOSBack started tracking a new policy.
It's the Automattic WordPress.com Terms Of Service.
June 2 around 2pm PT

DATA.GOV

TOSBack started tracking a new policy.
It's the Data.gov Privacy Policy.
June 2 around 12pm PT

DATA.GOV

TOSBack started tracking a new policy.

And this:

IMAGE·PERMANENCE·INSTITUTE

IPI MAIN PAGE | SITE MAP

The Archival Advisor

IPI's Guide for the Family Photo Collector, the Genealogist, and the Scrapbook Maker

HOME | ABOUT IPI | NEWS & EVENTS | MANUFACTURER SERVICES | CONTACT US

IMAGE GALLERIES
ARTICLES
BOOK REVIEWS
TIPS & TRICKS
FAQs
HANDY TOOLS
OTHER READING
NEWSLETTER BACK ISSUES
GLOSSARY
HELPFUL LINKS
OTHER IPI WEB SITES
SITE MAP

ARTICLES

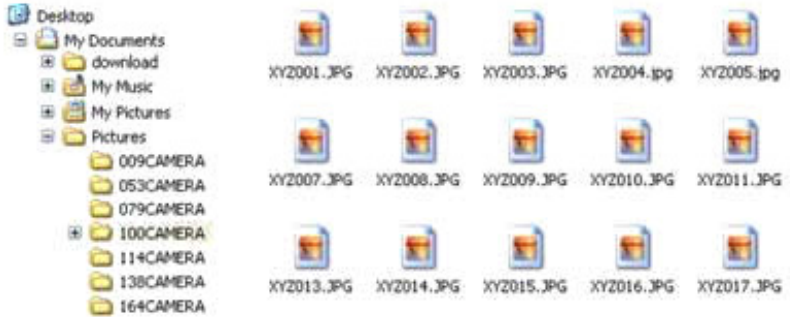
Preserving Digital Memory Files

Jennette Hanna and Daniel Burge

In the digital age, photographs are no longer captured as negatives. Digital cameras capture images directly to computer files, which are later moved via a memory card or cable connection to a computer. Traditional "analog" materials such as photographic negatives and existing prints can also be scanned to create digital image files. The advantage of digital image files is that they allow for easy editing of images, simplified copying, and electronic sharing. However, the goal of preservation will always be to ensure the long-term accessibility of the images.

Because digital image files do not require processing in a photo lab, and because camera memory cards can hold so many photos, the number of pictures being taken has dramatically increased. There is no longer a concern about the cost of film and processing, so users tend to take more pictures. They don't have to be as careful and methodical about which pictures they take. This has resulted in enormous collections of digital image files. And, because of these larger collections, users need to be organized so they can find the images they want, when they want them.

Organizing image files is particularly important, because scrolling through large numbers of directories – or worse, opening and visually checking files one by one – can take an enormous amount of time. Stacks of prints can be flipped through and sorted rather quickly. So the first step in organizing the files is to develop a file and directory naming system and consistently adhere to it. Most cameras assign a name to each picture file as it is taken – for example, IMG001, IMG002, and so on. However, that generic name provides no information as to what the picture is. Also, with some cameras, each time you put a new memory card in the camera, the counter goes back to IMG001 again. If you store all of your images in just one folder you may accidentally overwrite the older pictures with newly downloaded images.



Example of an unorganized file directory. The file and folder names are the default camera settings. It will be impossible find a specific image without opening all the files and folders individually.

If you have a common naming convention, such always using the date, event, and location (e.g., 2005_NYears_Chicago) to name a folder, then it will be easier find the picture you're looking for, and you will be less likely to overwrite or delete a file accidentally. Naming the folders in your directories by year, as shown in the figure below, will help limit the number of files stored in each folder, and it will make the folders easier to sort through. Ideally, each image file would be given a descriptive filename; however, this is labor-intensive.

http://www.archivaladvisor.org/shtml/art_presdigmem.shtml

Something to read (soon):

I, Digital: Personal Collections in the Digital Era

I'm editor, Society of American Archivists is publisher

Something to join:

Personal Digital Archives Working Group (PDAWG)

- Currently coming together
- Representatives from across the globe
- Focus:
  - Documentation & development of common toolsets for **curators** of personal digital collections
  - Documentation & development of common toolsets for **individuals** to manage & gain better control over their own personal digital collections
  - Collaborative authoring of several guidance documents
  - Engaging research communities that are most directly impacted as users or potential users of personal digital materials