

PERSONAL DIGITAL ARCHIVE CONFERENCE NOTES
FEB 16, 2010 – INTERNET ARCHIVE, SAN FRANCISCO
(REVISED VERSION, FEB 19, 2010)
Elspeth De Shaw

Speaker: Cliff Lynch

- Change in methods of communication/preservation – emails, personal networking, etc.
- Last 10 years of work in the field of personal digital archiving (PDA) have been descriptive work, understanding permanence/transience, applications
- Understanding “memory organization,” (libraries, museums) archives, etc. – what do those institutions want it to be?
- Building up personal collections of documents (to sell, share)
- Materials are becoming artifactual - entire computers, collections of media as opposed to scattered materials. Digital and physical are intermixed.
- Organize material as you create it (certain organizations are doing outreach to this end to educate the public)
- What is the appropriate role of archiving?
- Facilitating personal archiving and memory organization to serve our purposes - how personal memories connect to the industry of PDA
- Thinking beyond observation and analysis to policy, objectives, role of developments in our world

Speaker: Cathy Marshall (Powerpoint title: “3 things I’ve learned + 1”)

- 1) Benign neglect – most likely collection policy and stewardship technique (but not always a terrible thing)
- Stewardship is paved w/ good intentions (as opposed to action)
- Rare vinyl record anecdote: did not play rare vinyl record, archived it to reel-to-reel tape, now it is accessible on Amazon (saved the wrong thing, chose the wrong medium to save it on, now it is no longer important!)
- Growing amount of material (Flickr has 4.3b personal photos, Facebook 15-60b)
- benign neglect fits “what people do”: they move, don’t purge/cull, it’s easier to ignore than go through and organize
- 2) No single preservation policy will solve the problem!
- Two standard technical solutions: shove stuff into a database in the cloud to decode later OR safe storage and self-describing digital objects
- Knitting together storage and solutions
- Take into account things like dental xrays, HS yearbooks, catalogs, new institutions and cultural expectations (ie Facebook)
- 3) Forget about digital originals or reference copies, people store things in multiple places for storage – to share with different groups, etc.

- circularity of digital original (local copy as archival/highest fidelity), web copies have been augmented with organization and metadata
- each copy has a life of its own, they may look the same but are different (resolution, memory space, tags on photo, purpose of file)
- 4) given first 3 things, there will be interesting opportunities in searching and browsing
- Why it's different: re-encounter, faceted browsing, visualizations and desktop search, "the whoops factor"
- not "boiling the ocean": importance of starting one piece at a time and weaving together
- Felix Ungar or Oscar Madison approaches are typical
- multiple copies, each accretes metadata; challenges the idea of the "original"

Speaker: Gary Wright

- FamilySearch and Personal Archiving (FamilySearch.org)
- FamilySearch = a family history organization, non-profit (in Salt Lake City, UT)
- purpose: to link families and preserve heritage of mankind
- collects images of records from around the world (birth certificates, etc). Collecting 2.75 million images per week, annotated by 100,000 volunteers. 10 PB.
- preserves images on microfilm (2.4m roles of film), in Granite Mountain Records Vault
 - over 87 years, 3.1b images, more than 10b names (10% of all people ever on the planet), linking names into families
- Have technology to digitize microfilm collection, tidy up images (straighten, etc)
- collects new images from around the world, uses cameras & volunteers, and an indexing program to extract key info. (dates, names & relationships), transcribe about 1m names/day into a searchable database
- Two interpretations of each image to verify data
- Large-scale operation: thousands of users, millions of images, lots of partners worldwide
- Population growth and deep history are obstacles (no written records), paper-based genealogical records are slowly deteriorating (limited time frame to collect records)
- goal: facilitating access to digital records, developing techniques for digitizing physical records, sharing expertise and software, vendor relationships, volunteers, creating solutions to obstacles
- solutions to world's archives: digital research, provide computer access locally, links
- FamilySearch is investing to become a world leader in digital preservation (will have 13b objects by 2025), is consulting with key product vendors about scale, architecting to make automated tape the preferred archive medium (lower costs), also designing archive facility to protect \$2b collection of files
- Personal archives – valuable information held by family historians; FamilySearch digitizes material, links to BYU
- Biggest challenge is getting permission to duplicate (digital rights, personal rights)
- FamilySearch is considering a plan to allow patrons to upload items linked to specific ancestors (archival metadata will be required, formats will be restricted, initial forecast 50 pb by 2025); but this will require significant resources, etc
- Spent a week with a similar, large scope organization learning about their methods (...)

- Digital preservation is too difficult and too costly to tackle in-house, except for some rich organizations
- Everyone else needs: low cost cloud archiving services offered by providers on 3 continents (the EU won't store their digital files on a server in the US)
- costs will be driven down by provider economies of scale and using tape for archival storage. Low cost cloud services on three continents.
- growing by 1.5PB / year, 30 TB/week, 137.5mm images a year
- Personal archiving may be the key to making it work
- *Question*: will you consider storing genetic info in archives?? (no)
- *Question*: collection is not representative of communities/individuals with no physical archiving (ie oral archiving, local kinships)?? (FS is partnering with archives around the world; 70% of folks lived after 1700, and they are documented; Italy has lots of records in old churches, etc)
- *Question*: intellectual property rights/copyright - what about the ethics of gathering information from folks who don't want their records shared/accessed?? ("dead have no rights," but families have rights; make sure collectors understand what their rights are; can have restrictions on viewing/searching the materials; it's up to the archives to deal with legal ramifications of filesharing)

Speaker: Susan Thomas

- Personal Digital Archives at Bodley Library in Oxford
- working with PLANETS project
- Library first opened 1602; now has collections of manuscripts dating from 3rd c BCE, from all sorts of people (ie famous scientists, literary figures), and organizations, etc
- Problems: loss of collected material through media degradation and technological obsolescence; material we failed to collect
- Paradigm project: collaborative, holistic (archival lifecycle, from paper to digital); with contemporary politicians as creators; scientists as researchers (interdisciplinary effort); exploring cultural, legal and technological issues; addressing hybridity; JISC funded
- Case studies: 1) develop relationships with working politicians with a view to acquiring sample collections.... (more)
- Findings: Digital archive characteristics (scattered locations, public and private spaces, blend of personal and professional, simple formats but evidence of increasing complexity); creator's perspective (recent/digital = not archival, privacy concerns, importance of 3rd parties, trust, need for creators to become curators, level of commitment, "need to weed" personal v. professional)
- Metadata is **always** incomplete for some purposes. Metadata as an interface.
- Library's perspective: diversity of collection development models; need practical guidance and tools for capture/transfer; remain flexible while standardizing; digital repositories not ready for digital preservation; validate application of traditional archiving skills while still needing specialists for current issues

- Researcher's perspective: yet to grasp impact of digitizing; appear to trust the Library to provide authentic material; citation issues around migration, viewing objects in emulators; need more hands-on with digital materials
- Local outcomes: improve know-how in theory and practice; stand-alone digital archive; support for colleagues (...more)
- Many different types of archives; diverse media, formats, etc.
- Current project: developing a framework; changing processes; integration and hybridity; forensic analysis
- need "collection comprehension engine." Anchor is front end; malleable infrastructure with evolving services
- Future focus: interfaces and tools; digital preservation policy, acquisition (cloud/web archives), access (refining cataloguing techniques, etc)
- Areas of wider concern: clarifying disposal of cloud-based digital assets; more export facilities to get structured data and metadata out of cloud-based services; succession plans for cloud-based data; should cloud-base offer archive facility?; access to older software and license to use; improvements
- *Question*: hybridity: people provide access to personal digital archives in multiple places for different groups/services. With Library's collections, you develop strong connections with donators; donators give personal digital archives to multiple archival organizations... do you have complex relationships with other archival service organizations?? (shared services between institutions; depends on culture of each archival institution; the problem will change as time passes)

Speaker: Eric Saund

- Document Content analysis for digital archives
- Tasks supported operations for digital archives; all enabled by metadata, metadata layer and content layer, indexing items
- Have to get metadata from items; extract metadata from raw content (metadata is always incomplete, and is a static record; process needs to move to metadata as interface)
- automatic content analysis (ACA); refreshing data according to our needs
- Automatic content analysis rapidly evolving; analysis for audio, web, video, photo (who, what, where for each item)
- anyone who wants to build an archive should have access to automatic content analysis and metadata
- IE: Document analysis, separating text from images; extracting metadata; (we want a document we can understand in context)
- Category type (invoice, bill, itemized list, annotated); structural elements and relations (headers, logos, graphics, layout, handwritten annotations); relational context (construction project, supplier relationship, inventory and materials management): all relates to how we analyze documents and contents, gives us clues to "figuring them out"
- Academia (science based) and Industry (engineering-based) fields – for content analysis (not useful for personal digital archives), have different characteristics and different paying customers – as opposed to Hobbies: museums, schools, individuals, NGOs, etc

- Hobby projects: personal archives, physical archives and information – collecting and organizing
- Effortless document capture; scanning, camera-based document scanning/capture, etc.
- Collection comprehension engine: image processing to enhance images; OCR (character recognition); document structured modeling; classification, genre tagging, clustering; automatic cataloging; automatic document connection linking (duplicates/related documents); visualization GUI
- “The hobby stage brings together kindred spirits”
- *Question*: how to resolve issues of collection comprehension through time as technology evolves?? (anchor point = what does user want to do, what services can we offer?, infrastructure malleable enough to evolve with technology and changes in collection types)
- *Question (unheard)??* (scene-text analysis application (Google); point digital camera at text and have it automatically recognized). *Question*: How do I get ahold of it? (Go to paper, Japanese camera vendors)

Speaker: Bill Janssen

- Managing physical and digital archives; keeping track of it all
- Use browsing tools, applications
- System that guards a set of documents; acts as web service (most recently used documents: is searchable)
- projections are formed in a client and sent to a server, creates simplified projections so they stay readable (using plain text, imaging) through time
- separate document into metadata, images, text, wordboxes, links (stored in simple files)
- after projection: text version of document, page images, basic metadata, wordboxes – individually searchable, easy to see what’s going on in the document
- Then goes to document analysis engines; adds more metadata, stores results
- After ripping, there is even more information and more metadata, document icons, page thumbnails, etc.
- Server is guarding ALL the data
- Save as for IMAP or Web server
- Uplib has a Memory Palace
- Repository structure: repository, document subdirectory, document folder, page images/thumbnails/html/originals
- ReadUp document reader, java widget; works on any document format, can browse entire personal digital archive collections
- Future proofing: use projections in common simple format, documented; will make it easy to add parsers for new document formats; make it easy to modify server; use peer-to-peer instead of centralized service to scale up and network repositories; is easier to manage (not Flickr or IA, maybe CCN), makes every file accessible regardless of format (web, cache, etc)
- *Question*: CCN?? Center Content Network (??); (content sharing – accessible to whoever, wherever; makes sure documents never get lost because they are saved all over the place)
- *Question*: enabling availability?? (Just need to know name of doc)

Speaker: Nancy Van House

- Personal photography
- We care about photos because of: historical importance; they have significance for larger world; contribute to studying material culture (what life was like in history); have larger documentary significance (ie: contributions to biodiversity research, environmental research affected by photos of plant life); we can see when history is being manipulated (forensic photography, altering photos to influence records)
- Print photos tend to be contextualized; we get meaning out of context and annotations; ie relationships, locations, dates, stories
- We lose information when we scan images: Embossed information, pencil annotations etc: lose context of document
- Information on postcards/slides – problems reading other people’s handwriting, annotations in different languages, etc (evidence might be lost by scanning)
- What happens when we go digital?
- Tons of identical photos; little changing, little deleting duplicates, etc; little metadata; tons of meaningless file names
- People don’t think their archives will outlive them – they won’t go anywhere
- Lycra: professional photographers talking about metadata; collects information and makes processes easier (keywords, titles, copyrighting, coding, geolocates by date and time of images)
- It has 2 files for every image; sidecar file (metadata with all changes, AND an original)
- Not all the same file extensions; standardizing digital negative files makes for tons of images; personal photographers deal w/ these issues
- *Question:* type managers – way around type managers in the future?? (I don’t know! Anxiety about preservation of images because of formatting issues; most folks concerned with current issues, not with future problems)
- *Question:* how are you managing your harddrives?? (have lots of copies of everything, also need to have stuff accessible off harddrives, outside of the house)
- *Question:* is it important to make selections?? (some people make selections); *Question:* what would you teach people about making selections beside deleting duplicates?? (indexing = predicting future use; emotional significance of images)

Speaker: Heather Champ

- Flickr: 4.3b photos, 3.5m new photos/videos every day – tsunami of items
- Account creation around life-changing events (births, deaths, weddings)
- Personal stories told through pictures, issues around picture sharing (intimate moments, public v. private)
- Also documents large historical events: communities respond to these images

- 5000 photos uploaded per minute: difficult for people to search documents, navigate privacy levels and metadata (tagging, description, location)
- What happens to personal photos in a public realm? (unrelated people start adding information about the photos – historical, cultural – turns into collaborative activity)
- “I found a camera” situation: posting pictures to find owner of camera (people use clues in photo to identify it and figure out context)
- Stories of photos can change over time: placing small, detailed photos into larger context (ie: hummingbird feeder in window; eventually integrated into view of street (and then geolocated))
- Physical photo albums live in a safe; digital photos live in a Flickr account (or the camera becomes the photo album);
- one person having 4 flickr accounts: publishing personal, life-changing events; transition from dealing with photo albums/privacy to digital images/public
- Tip: include the usernames and passwords to all your social networking sites in your will
- *Question*: Storage of files after death of user?? (If family can provide death certificate they can access/delete account; no activity after 90 days leads to automatic deletion)
- *Question*: Issues of scale, amount of material over time?? (Flickr makes 6 copies of each photos; Yahoo gives money to grow capacity of storage over 4 “arms”;
- *Question*: cost?? (No idea)

Speaker: Snowden Becker

- Center for Home Movies: collecting/preserving/providing access to and promoting multi-disciplinary approach to amateur film and video
- Home Movie Day: education and outreach about film, preservation of individual works, partnerships with local groups/organizations; thinking of documents as community history: documenting ways of life, material culture; inspect and project films (happens on every continent except Antarctica)
- Risk to materials in good condition in clean equipment run by qualified individuals is LESS than risk of continued ignorance/transferring to different formats, easier to grapple with formats of moving digital material
- solving problems of materials with short “shelf life;” problems of metadata
- Films AS home movies (not documentary); represent great depth of culture, exceptional footage in each and every home movie
- Preserving individual films; presenting them to the public, making them more accessible
- Stop thinking about film as family-specific; broadening view of movies in cultural context (others find meaning in them, can analyze films; different views/perspective on home movies from childhood to adulthood)
- Home video as a venue for learning about audio/visual materials: talking about it raises awareness and brings materials to light; reconsidering materials (not just home movies, but all media; if you build it/prepare for it, it will come out of the closet); condition of materials is good (domestic environment is conducive to preservation); also, materials can be too expensive to upkeep, etc; identification of what is important, assuming responsibility for maintenance of and access to digital materials

- Implications for the future: records lifecycle is shortening, so preservation must be a recognized need; archivists are inadequately prepared to work with mixed materials; many “mousetraps” may be improved (technical and technological; managing inheritance and ownership transitions; rights, etc)
- Future: Creating a corpus of accessible material, enabling and encouraging virtual connections (networking, maintaining, altering how we view digital materials)
- Establishing “godparent” role for archives – kits for preparing a preservation plan, “wills” for disposition of media materials
- *Question: Date?? (9/23)*
- *Question: web archives hosting/operating digitizing, building workflow, accessing personal home movies to digitize?? (matter of finding stuff that’s available; outsourcing work of describing/analyzing documents)*
- *Question: piles of media stacked up – dealing with time issue?? (Stays independent of accumulating media; some people collect media; analyze and index more efficiently; accumulating lots of media and analyzing it all)*

Speaker: Marc Smith

- Data structure that needs to be archived: NETWORKS themselves (not just data IN networks)
- Toolkit called “NodeXL” to preserve/analyze social media networks
- Set of graphs that allow the rest of us to browse the network itself
- Building tools for users to manipulate and analyze graphs
- Wants metadata as it relates to social network theory; analyze sociological aspects of connections on the web
- “Sociometry;” positions and relationships of people in a social “graph”, can then apply this theory to social networking media; mapping newsgroup social ties
- Distinguishing attributes of online social roles; answer person (outward ties to isolates); reply magnet (has inward ties from local isolates)
- NodeXL (Network Overview, Discovery and Exploration for Excel): filtering, sorting, organizing graphs of dynamics of social networking sites
- Tweeters: following, replying, who does and does not follow each other; online cliques (NodeXL can analyze who is the bridge, who is in which clique/cluster, why are(n’t) they connected cross-clusterwise) (can also be used to simplify voting records of Congress, for example)
- Mapping from what is on the network to the computer’s display (looks like Excel sheet)
- High “degree” vs. high “betweenness” (influence vs networking/gateway)
- Can sort by subject (who mentioned “SAP” on Twitter, how are they connected, who has influence in this group, etc); Graph analysis (of clusters, etc)
- Continuing process: ongoing collection, additional sources, more metrics, clustering, cross-platform, etc.
- *Question: how many rows in XL?? (Millions)*
- *Question (unheard)?? (graphing in XL format; archiving the graphs)*
- *Question: template for office archiving ?? (working on it; Node for open calc)*

Speaker: Ben Gross (Powerpoint title: Archiving Identity)

- Identity and identifiers (what we are/know/have/are assigned)
- Identity and personality (are not the same thing), multiple identities different than multiple personalities
- segmentation and integration of identities and identifiers
- How many emails/IM ids/virtual characters/social network profiles/phone numbers does each person have?
- More than one identity, for social/political/technical reasons: there is a continuum of segmentation
- What people *have* VS what they *want* VS what they *have time for* (have multiple usernames, passwords, accounts – implications about archives and access to archives)
- When do these services etc “go away”(data evaporation)? Days without login leads to deletion of account (30-270 days, depending on site), recycling of domain name; accounts/files permanently deleted and permanently inaccessible
- Data evaporation after death for personal accounts: do sites make a memorial, still allow access, or simply delete?
- “Information economy: all activities leave traces/fragments that can be retrieved and amplified.” One identifier gives access to others (leads to security problems) – if you can access one account, you can reset all others
- Re-identification; sifting through and correlating data; integrating facets of identity and identifiers
- How do identifiers play into archiving? There are facets of our lives in each item – where do they come from and why??

Speaker: Cal Lee (spoke quickly - can search for “Cal Lee UNC” to retrieve Powerpoint)

- Traces convey information: we can select traces, create richer traces, make extra copies
- collectors maintain traces, capture, etc (?)
- resources are limited, and collecting is expensive
- advance curation of digital archives (“get, grab, guide” concept)
- extract useful information before copies become obsolete
- selection decisions: which data on disks, in word documents, etc
- Grab from web; maintain accessibility
- Guide: professionals only have responsibility for a tiny sliver of archived docs
- Data liberation front (getting information/items off web, rights to accessibility/privacy)

Speaker: Francesco Spagnolo

- Magnes Museum: artifacts from Jewish diaspora (all kinds of formats and media)
- Trying to format digitization
- Narrative theory: preservation IS a representation; selection, history is told by what is included AND by what is excluded
- examining online presence and narrative: decomposing it and fusing to recompose more complete comprehensive narrative
- Digital narratives, using a memory lab: users can bring memorabilia to scan, then create narrative (finding a voice, voicing a community), relates to oral history and fieldwork
- integrating 3 venues for presentation online; databases (social media, etc) analyzing how these 3 venues interact
- Flickr is not just about photos, but is about all images: put digital items on Flickr from museum, started composing stories
- only a few slides selected to create the narrative (each narrative has a beginning and an end) – slides are individual fragments, allow for easy access by public, create a “public face”
- Memory Miner: database about images with two outlets: uploads to Flickr, and creates an html interactive document which is searchable, can associate with different urls, (external audio files, etc)
- Same content gets uploaded to Flickr, in same sequence, with same titles, etc
- Concerns of cultural institutions = entering “web2.0” and losing institutional voice
- Presences also linked back to collection database
- Paradigm of “dusty archives” resists realities of digital media age
- Each online venue links back to each other, are interconnected; each acquires unique connections through its own venue/audience
- People bringing in their personal materials; create their own archive in the way that they want, make their own narrative
- Memory Miner at the Magnes Museum (see YouTube video about software usage)

Speaker: Dave Marvit

- archiving “lifelogging”: a selection process, which has implications for narrative: if nothing is excluded, what is the nature of the narrative?
- we all try to narrativize our own lives to assert our own importance
- everyone puts themselves at the center of their own narrative PUBLICLY via social networking sites
- issues of public and private, boundaries between these two realms
- transformational force comparable to AIDS epidemic breaking boundaries about sex; opened a gigantic can of public/private worms: questions about what is appropriate to share on Facebook, Twitter etc.
- *Question:* what is the impact of AIDS epidemic on public consciousness, as opposed to feminism?? (response unheard)

PANEL

- Can't "select all" of personal digital archives; narratives are made of fragments. The whole is missing (parts are suppressed, etc); certain humans don't fit within the parameters (of having PDAs, access to them, etc), and therefore are excluded or invisible from narrative: social networking brings these populations back into the public sphere
- people use these spaces (social networking) for selective disclosure; selective decisions are restricted by capacities of sites/venues (ie, search queries on youtube, ranking relevance in a search), etc. – this is confusing
- metacollection is biased/misrepresented – biases towards things that are easy to collect (especially in collecting institutions)
- *Question*: introduction of children into discussion – kids on Facebook, Twitter, etc; exclusivity of media collecting (who has the resources to do digital archiving, past and present?) ?? (Digital environments hard to "flood," can store all the phone videos taken by kids, etc...)
- which kids have access to YouTube, to digital recording, etc across the globe? Even though we're "flooded," there is a huge population that is not represented at all. Some cultures don't even talk about the deceased (taboo); how do we help these folks find a voice? (anthropologists, etc.); fragments of narrative – what is around it and what is missing from the picture?
- *Question*: content that we create and choose to share; what about connected devices? What about involuntary traces (ie, on GPS things, etc), or items that are not produced consciously?? (Fundamental argument: traces change monumentally when imbued with meaning/are symbolically representative; what pieces are used to construct the narrative?? How do we assign meaning to the traces we leave?)
- history – personal, family, communal; managing the fragmenting nature of narratives; how many voices are intersecting (consciously or not)?;
- how do individuals talk about themselves; how do they express themselves to others?;
- success of narratives through time – what makes them successful?
- what would a narrative of unconscious fragments (ie mpg trackers or GPS location devices) look like?
- dusty archives reference – dust signals low accessibility and/or low importance
- what is important VS. what is absurd?... (ie the same photo is described as completely different things in different contexts/venues)
- History forgets – narratives are about forgetting AND remembering... what is forgotten? is it chosen to be forgotten? why?

Speaker: John Kunze (Powerpoint title: Preservation is not a Place)

- Curation design principles: preservation is not a repository; it IS an outcome, a relay; all collections need some aspects of preservation; permanent objects live in disposable systems (need to be easy to escape from)

- Microservices approach: low barrier, low commitment tools; avoid monolithic, single-culture systems; compose repositories from small, independent, interoperable microservices
- wisdom of files? ; modern filesystems are fast, stable, interoperable; native OS tools create, list, change and backup; therefore file-based microservices will be easy
- Collection storage: Pairtree (use pairs of an object's identifier characters to create the object's file system path)
- Pairtree storage service creates paths from objects identifiers, resulting directory collection holds objects of any type
- Import a Pairtree and you can enumerate all objects and identifiers; produce any object by requested identifiers; maintain and backup tree with ordinary OS tools; rebuild broken catalogue by "walking the file system"
- Files and filegroups (directories) are important: directory level metadata revealed via filenames (Name-As-Text "Namaste" Tags)
- Object storage: Dflat – a "digital flat"; a residence for object data and metadata
- "File level reverse delta compression;" simple, effective
- Making metadata simpler: can be human and machine-manipulable; email header protocols drove early internet, so why not metadata? (ANVL; A Name Value Language)
- Simpler metadata semantics: ask not what metadata you would like, but how little metadata you need NOT to screw up bit-level presentation
- Dublin core is both too hard and too soft; requires who/what/where/when, or explicit reasons for absences; comes with author name and date conventions; accommodates very rich and very poor metadata
- Simpler tabular data: scientific data commonly recorded in Excel (not good at versioning, schemas, date formats, etc); what add-ins could be defined to make research data more sharable, publishable, archivable?
- Success story: BagIt. Library of Congress uses it to receive grant-funded partner content; file packages suitable for disk-based or fast network-based transfer or generic content
- what has already failed? can we generalize about approaches that are likely to fail over time?

Speaker: David Rice

- Embedded metadata vs external metadata (external = card catalogs, databases, finding aids; helps you find the actual data)
- Securing relationship between metadata and essence: what is the content? where did it come from? how can I get more info? what rights do I have to this content? (email and photos, get separated quickly); metadata gives key pertinent info about files
- Metadata 1 – embedded in a webpage, searchable; Metadata 2 - embedded within file, not searchable but exchangeable; Metadata 3 – derivative of embedded metadata, facilitates collection management and local search and browse (ie: form => webpage => download => file player); embedded metadata makes files easier to search/manage
- Different standards for embedding metadata; BWF MetaEdit: tool that allows for embedding, editing, adjusting of metadata

- Rules followed or not followed in industries/organizations (...)
- DV Analyzer; error detection and quality control; DV Analyzer analyzes DV video files and reports on: video error concealment, audio error codes, structural incoherencies, timecode inconsistencies (blurriness, damaged tapes, etc) – ie; makes substitutions from prior frames to fill in damaged parts of frame
- Temporal Metadata Analyzer; analyzes video frames; jumps in time; jumps in frame, etc
- AVPS Project – Faceted technical metadata aggregator project: efficiently and accurately document technical metadata of a large file set of digital video; identify trends in online digital media; compile data that can inform obsolescence analysis; detect use of embedded metadata (by user, hardware or software)
- Downloading files from web is time consuming; downloading relevant files about metadata, not actual content file
- difficulty of downloading/searching certain files (misspelling, etc);
- embedded metadata trends and challenges; video in wide variety of wrappers, etc; transcoding utilities are less metadata aware than image transcoding utilities; less social expectation to embed metadata into distributed videos compared to music.....etc
- *Question??* technical analysis; pdf files are wrong a lot – having standards for metadata

Speaker: James Jacobs

- preserving our digital heritage; the community taking control - LOCKSS (lots of copies keeps stuff safe)
- Libraries use LOCKSS as digital stacks (local custody and preservation of web-published collections); publishers ensuring their content remains available
- Since 1998 at Stanford U, self-supporting since 2004
- Two kinds of implementation: public LOCKSS network (libraries/publishers), and private LOCKSS networks
- Distributed digital preservation system – open source peer-to-peer software
- Authoritative versions of heterogeneous content (journals, books etc)
- Public: 400 publishers, “general library collections”; private (theses and dissertations, images, etc)
- How does it work? Library installs LOCKSS software, publishers give permission; need a machine with enough CPU and memory for software (LOCKSS is like a “shelf”)
- Independent collection: LOCKSS box puts content into caches; software audits and repairs all of content in network; creates an authoritative copy which is distributed to all networks, maintained over time through continuous audit and repair
- decentralizing preservation, makes for minimized damage
- Private networks: a community, sharing preservation responsibilities (ie digitized images, ejournals, ebooks, theses, government documents); statewide government networks (Alabama), can be digital preservation for social sciences; public records, etc.
- Private LOCKSS networks interoperate
- Locally focused special collections and archives; local history/museums, local content
- *Question:* Personal libraries?? (libraries might help facilitate this; they have institutional knowledge to create metadata, make space on their own servers for individual content)

- *Question??* Zotero as collection tool, not just citation tool
- *Question??* culling down to one place, VS importance of LOCKSS concept

PROJECTS

Speaker: Cal Lee

- Digitalcurationexchange.org : cluster people around common interests

Speaker: Will Snow

- SALT project – 2007-2009 grant project; Stanford U: Self Archiving Legacy Toolkit; takes documents and normalizes them into standard pdf format; run through web service to extract metadata; create faceted browser view to connect dots in archives; finalize and present on web as personal legacy
- “Snap to grid”; edit tools; for luminary, archivist, legal = faceted browser to find documents easily (keywords)
- visualization of legacy (based on metadata) – timelines, organized by project, etc. = visualization leads to stories!
- SALT results – scanning = expensive; structure from digital and online; ran into physical limitations; audio annotations; grouping, rating, keywords; “way too industrial,” – can see what the internet “says” about a person
- Next phase: online visual representation of the career and legacy of 2000 faculty/researchers
- automatic creation, navigating the draft, having influence/collaboration; Project fit (funding, faculty network); Collaboration; Total Recall

Speaker: Tom Munneke

- conversions between two (... unheard)
- online healthcare and personal health records – personal archives
- 75 % of medical information is metadata that points to nothing; metadata becoming data (metametadata)
- Medical record technology has funding, shifting from enterprise to personal members of medicine
- personal archiving; genetics; 18 family members have genomes (can document hereditary problems, etc); diachronic information systems (synchronic = snapshot); flow of information over time
- role of language in archiving: changing names of same disorder; creates problems for archiving
- Recording over time – logistics of HOW to do it?; how to deliver items to future recipients? (futureme.org – sending letters to future recipients)
- Pendix (pending index); building a future narrative; transactional model detects things that happen; creating a future common space
- Wiki: a historical log of what’s happening; binding to a future state
- building things towards a future: stories, expectations, etc....

- Ancestor principle workshop; are we being good ancestors? looking forward, looking to history to inform our decisions about the future
- personal narratives have power to move us forward, not just organizing from the past

Speaker: Bruce and Sue Wilcox

- Speaker for the Dead
- Digital recording and search of all personal digital archives, digital immortality
- websites that store all your “stuff,” are just other places to store things (like a digital “shoebox in the attic”)
- how can experiences take on a life of their own?
- concept from science fiction book: curator for memorabilia from your life, a speaker for the dead: database of collective memorabilia; presents your life via a 3d avatar – converses with you as though human!
- “Chatbots” – used to represent major companies around the world (ie IKEA online helper)
- live on terraformed planet... replacing PEOPLE in an artificial, online reality
- creating digital speakers for famous scientists = virtual scientists
- Examples: Suzette; chatbot; can store lots of personal digital archives; 3d virtual cities application (from Apple??)
- prerecorded pattern matching ability, content; chat language matches patterns of MEANING, not just patterns of words (space and negative space); synonym sets; based on synthesizing information, not just rules of grammar etc; analyzing phrases from multiple perspectives, synthesizes response; pattern matching...
- Dead folks, alzheimers folks, your own avatar could replace you online

Speaker: Judith Zissman

- “shoemaker’s children wander around barefoot”
- what is our responsibility as attenders of an event to archive our own experience? what do we want to remember and share?
- collective and event-based personal archiving present many of the same questions we ask about archiving in general: who is the audience? who accesses? what is the lifespan of the information? are we designing for conversation? tools for real-time broadcast vs longterm records? rights issues? do we want to differentiate personal and professional involvement (general searches don’t filter this)? do same “best practices” apply equally to SXSW and the European Science Foundation Conference on Functional Genomics and Disease?
- how do we preserve what we’ve talked about today?

Speaker: Jeff Ubois

- conversations surrounding the generating of this very conference
- David Rosenthal paper: cost of “petabyte” data? cost of “terabyte” data? fixed amount
- data liberation front; using tools to escape from social networking; data accumulating outside our control
- client uploaders; software uploads to social networking sites; need a box to check to send a copy to different venues (archival sites, etc)
- broadcast archives around interface design; competition for interface design

- social networking from 1950; citation analysis from 1950 articles, etc; using this one man's collection as a test set for archival techniques (interface design)
- terms of service; what to do when service cannot guarantee survival of stuff: standardizing, etc.
- associating what one has studied with physical architecture
- donor agreements, standardizing – privacy, preserving if you provide metadata, need to be translated into digital world
- types of data: financial, healthcare records etc.
- Request: looking for ideas around guidance for libraries, museums and archives; what can they do? Custodians of their national heritage; examples of good policies in these areas

COLLOQUIUM

What problems are worth investigating?

- funding and interest (by a company, individual, etc?), what are the costs and how will the costs decrease (advancing technology, etc)? where does the money come from – how do we just get this done? (collecting statistics)
- personal: narcissism is very important to archiving; a collective of people becoming aware of issues; educating people on what the consequences are; raising awareness
- expectations of our personal documents to survive
- Gordon Bell wants to have records totally in cyberspace (not in a museum with artifacts); integrated circuit; digital taxonomy of content; information that is irrelevant to content; working group to recommend how to digitize content (can't just hand someone a harddrive); encouraging institutions to take on the problems;
www.mycybertwin.com/gordonbell ;
- hybrid archives – just a blip in time?; saving all of Facebook to keep a record; when will photos on Flickr be taken down?
- personal choice about what others post/make public about you; privacy??
- talking about a 2-4 yr experience vs a longterm record; what is useless and boring? why do we want to digitize everything? what purpose is it serving? common frameworks; once you have a tool and can repeat it, the tool will continuously improve
- talks we can have in one year's time to continue the dialogue? what services derive from data? what can you do with it? can you construct a narrative? what conclusions can you draw from the data? from personal to small organizations?