Future-Proofing Digital Records Management: Emerging Concepts from Current Research

Patricia Galloway
School of Information
University of Texas at Austin

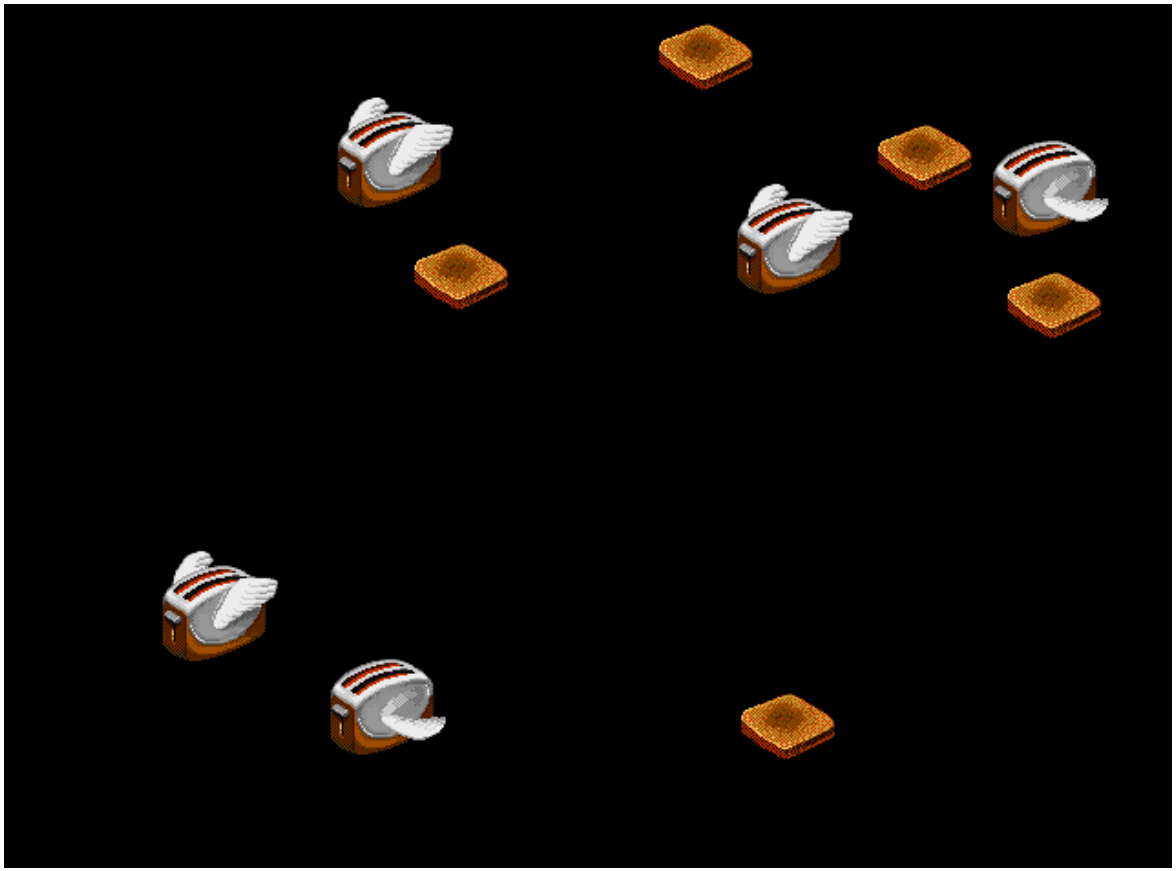
Infrastructure History

- Mainframes + terminals
- Mainframes + PCs
- Mainframes + diskless PCs [NOT]
- Mainframes + PC networks-on-servers
- Mainframes + networks + cloud(s) + mobile devices (digital appliances) + specialized workstations

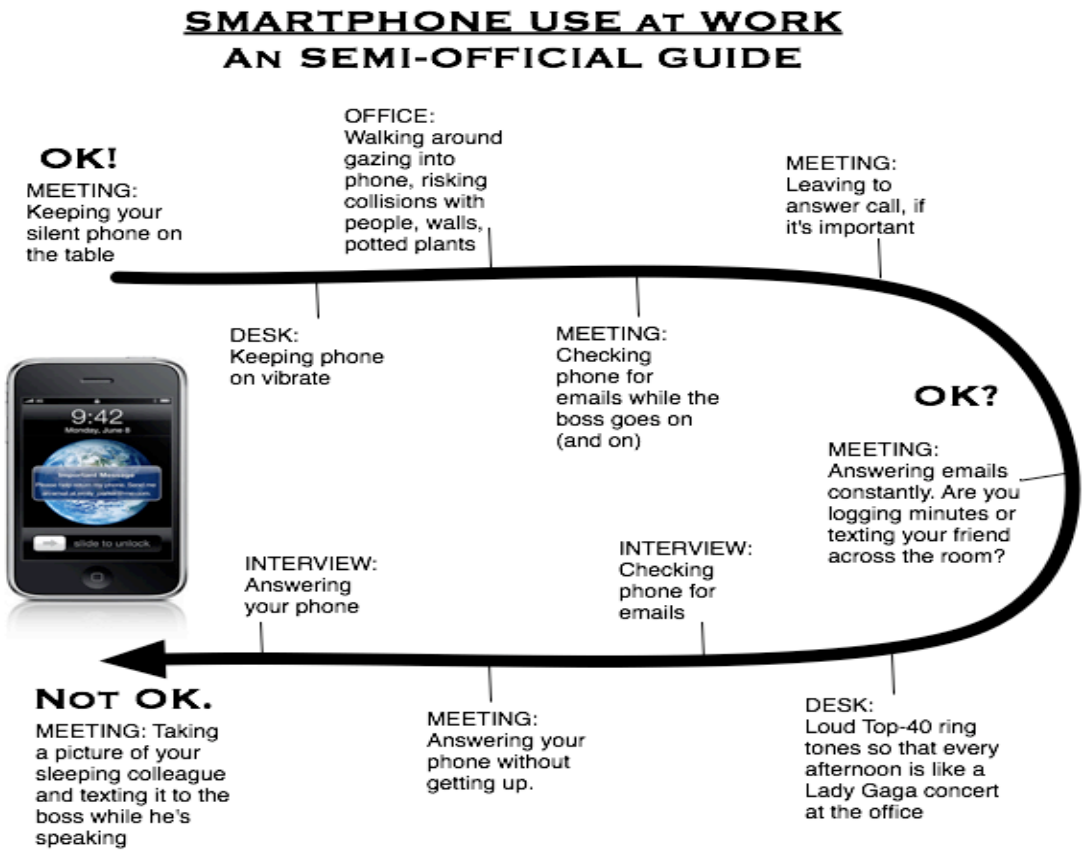
Cloud-cuckoo-land

- Mainframe = cloud
- Shared systems on servers = clouds managed across systems
- External SAAS you don’t provide (e.g. social media, email) = cloud (unmanaged)
- Internet access = route to many clouds (unmanaged)
- Is centralization the only answer?
- Will centralization even work?
- Well, maybe: on different scales

The future of personal technology



The future of personal use at work



What are digital records problems from a policy perspective?

- Security, privacy, confidentiality
- Realizing the promises of automation for work within the organization and service to the citizen
- Information asset preservation for current use, for compliance, for historical archives

Refactor digital records problems to RM/archives perspective

- Dispersed (agency-restricted) need for management of e-records in the short term: statutory, policy, procedural constraints vs increasing demand for flexible use of new tools by employees
- Short-term uncertainty of the value of records when the affordances of ever-new formats and applications are unknown (in the digital environment, permanent retention can and often must *evolve*)
- Long-term need for management of information assets once we know they are assets: *knowledge management* that may kick in at any time

How real are digital records problems?

- Technological problems are real but yielding to engineering solutions
- Social problems are much more intractable
  - Policies and procedures require understanding of technology and willingness to deal with change
  - Employees are more computer savvy and want to use accustomed practices
- Sustainability will be a serious issue as long as renewable electric power is not assured

Digital archives research from universities (some examples)

- Theoretical study of archives as representation
- Practical techniques for trustworthy capture: digital forensics
- Practical techniques for preservation
  - Development of digital repositories
  - Standardization of preservation metadata sets for different digital media

Solutions from government and business

- Best Practices Exchange http://www.bpexchange.org/
- Open-source projects
  - Linux (OS)
  - MySQL, PostGres (RDBMS)
  - Apache (webserver)
  - Sendmail (email)
  - Lucene (search engine)
  - Open Office (office suite)
- Microsoft’s -x formats
- Adobe PDF-A

Solutions from global initiatives

- Digital repository management
  - DRAMBORA (audit)
  - TRAC (audit and certification)
  - PLANETS (repository planning)
- Format documentation: UDFR + DROID
- Preservation metadata: PREMIS
- Internet harvesting and preservation tools: http://netpreserve.org/software/downloads.php

Digitized materials management models

- Digital libraries
  - Greenstone
  - ContentDM (OCLC)
  - Omeka
  - Hathi Trust http://www.hathitrust.org/
- Digital repositories
  - DSpace
  - Fedora
  - DAITSS

Born-digital materials management models

- PEDALS http://www/pedalspreservation.org
- Internet Archive http://www/archive.org
- ICPSR http://www.icpsr.umich.edu/icpsrweb/ICPSR
- Portico http://www/portico/org/digital-preservation/

Some generic solutions

- “Trusted repository” concept and tools
- Centralized/LOCKSS-style “clouds” providing secure storage, offsite and far offsite but controlled by owner
- Generalized and privileged authentication via improved and simplified methods (shibboleth, public key infrastructures)
- Escrow archiving and the necessity of digital records management moving beyond static standards
- Global management of format and migration tool repositories, which permits:
  - Acceptance of all formats
  - Migration on demand
- Centralized management and capture of email, coming, going, and internal

Will someone else do it? Maybe…

- Content (so far)
  - WWW (some deep web): Internet Archive
  - Twitter: Library of Congress
- Infrastructure
  - File formats: UDFR + DROID (2012)
  - Old code: Internet Archive, computer museums
  - File viewers: Not yet being pursued formally
  - Emulators: informal alliances with gamers, work by IBM

Some things we know now

- All recorded information needs management and evaluation
- Establish management before records are created, but don’t let it petrify: must be extensible, scalable, prepared to embrace change
- Establish automated controls to apply metadata, track status, ensure authenticity: granular employee action won’t scale, not even tagging
- The 5% solution is a dead model for digital records: it also won’t scale, and large corpora of records represent business process better
- Appraisal will be done with digital tools so we know what we have instead of guessing
- Mass conversion of formats is a dead end; migration-on-demand will be the solution

Three-tier model (traditional records, local management)

Records lifecycle	Active use	Semi-active use	Information asset
Host	Local agency IT	Trusted private cloud	Archives
Evaluation of use	Provisional schedule	Long-term appraisal	Reappraisal

Tier 1: Records lifecycle

- Generation and active use: observe use patterns by home agency, other agencies, and public for preliminary valuation
- Semi-active use: observe use patterns by home agency, other agencies, and public with access
- “Inactive”: observe use patterns by all agencies and public with access
- Afterlife: observe use patterns for ongoing evaluation

Tier 2: Evaluation of use

- 1. Provisional schedule developed when new series is created and enters active use: use digital discovery tools
- 2. Long-term decision might dictate sending to dark archive after period of semi-active use: evaluate usage statistics, holds, public input
- 3. Reappraisal of inactive records before formal archiving; then reappraisal every ten years

Tier 3: Hosting = access control set

- Local agency IT structure
- Trusted (private) cloud
- Archives (semi-dark; escrow)
- Archives (open)

Web 2.0 for official records?

- Dynamic websites backed by DAMS
  - Database preservation problem (backend preservable)
  - Web-based materials delivered from database
  - User interaction captured to database
  - Versioning must be tracked
- Social media used for official messages
  - External third party mediates messages unless private cloud
  - Messages and histories must be documented
  - Downloads possible? (cf. LoC and Twitter)
  - No control over changes in software
  - Best solution may be capture at time of posting

Web 2.0 in three-tier model

Records lifecycle	Active use	Semi-active use	Information asset
Host	Third party commercial (private?)	Third party commercial (private?)	Archives (extraction?)
Evaluation of use	Access to analytics?	Access to analytics?	Appraisal? Reappraisal?

How might practical digital records archiving work?

- In most states, a secure and trustworthy archival cloud environment could be built and managed by the archives, IT function, and Attorney General, as a service to agencies for digital records, parallel to that of a records center for paper
- Sophisticated authentication methods could mirror local access to these records for agencies
- Noncurrent records could be escrowed into this environment; agencies could retain gatekeeping for public access and full read-only access for themselves, operators would guarantee authenticity
- At disposition, archives would take formal ownership

No time like the present

- The digital records environment will continue changing (e.g. static to dynamic websites)
- Almost all records are created digitally today, and most are communicated that way, hence they <u>are</u> the records: paper is the use copy
- Good examples are now to be found widely, good advice is being discussed globally, and leading institutions are taking on infrastructural tasks
- Good enough digital archiving is better than lost records or another warehouse full of paper

Questions?

- Contact information:
  - Patricia Galloway
  - School of Information
  - University of Texas at Austin
  - galloway@ischool.utexas.edu