Getting ETDs off the Calf-Path:

Digital Preservation Readiness for Growing ETD Collections and Distributed Preservation Networks

Gail McMillan (Director, Digital Library and Archives, VA Tech)
Martin Halbert (President, MetaArchive Cooperative)

ETD 2009 Meeting
Pittsburgh, PA
Thursday, June 11, 2009

What is the Calf Path Syndrome?

One day, through the primeval wood,
A calf walked home, as good calves should;
But made a trail all bent askew,
A crooked trail as all calves do.
Since then two hundred years have fled,
And, I infer, the calf is dead.
But still he left behind his trail,
And thereby hangs my moral tale…

...The years passed on in swiftness fleet,
The road became a village street;
And this, before men were aware,
A city's crowded thoroughfare;
And soon the central street was this,
Of a renowned metropolis;
And men two centuries and a half,
Trod the footsteps of that calf.
Each day a hundred thousand rout,
Followed the zigzag calf about;
And o'er his crooked journey went,
The traffic of a continent.

A hundred thousand men were led,
By one calf near three centuries dead.
They followed still his crooked way,
And lost one hundred years a day;
For thus such reverence is lent,
To well-established precedent.

-Sam Walter Foss,
“The Calf-Path” 1896

Overview

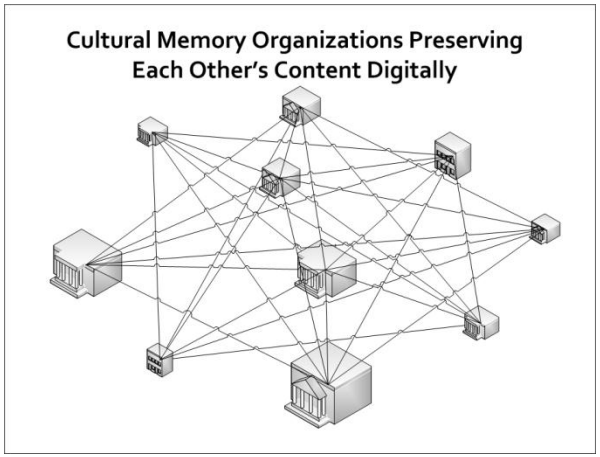
- This paper will address **"Calf-Path" problems for ETD repositories**, problems associated with the early ad hoc and idiosyncratic workflow patterns
- This paper will document relatively simple **principles and guidelines for such ETD programs** that can greatly improve the subsequent likelihood of implementing successful distributed digital preservation programs
- These best practices **can benefit start-up programs** that have not yet established regular procedures and standards for directory structures, metadata, and file naming conventions
- These guidelines were **distilled from workflow analyses within the MetaArchive Cooperative**

MetaArchive Cooperative Overview:

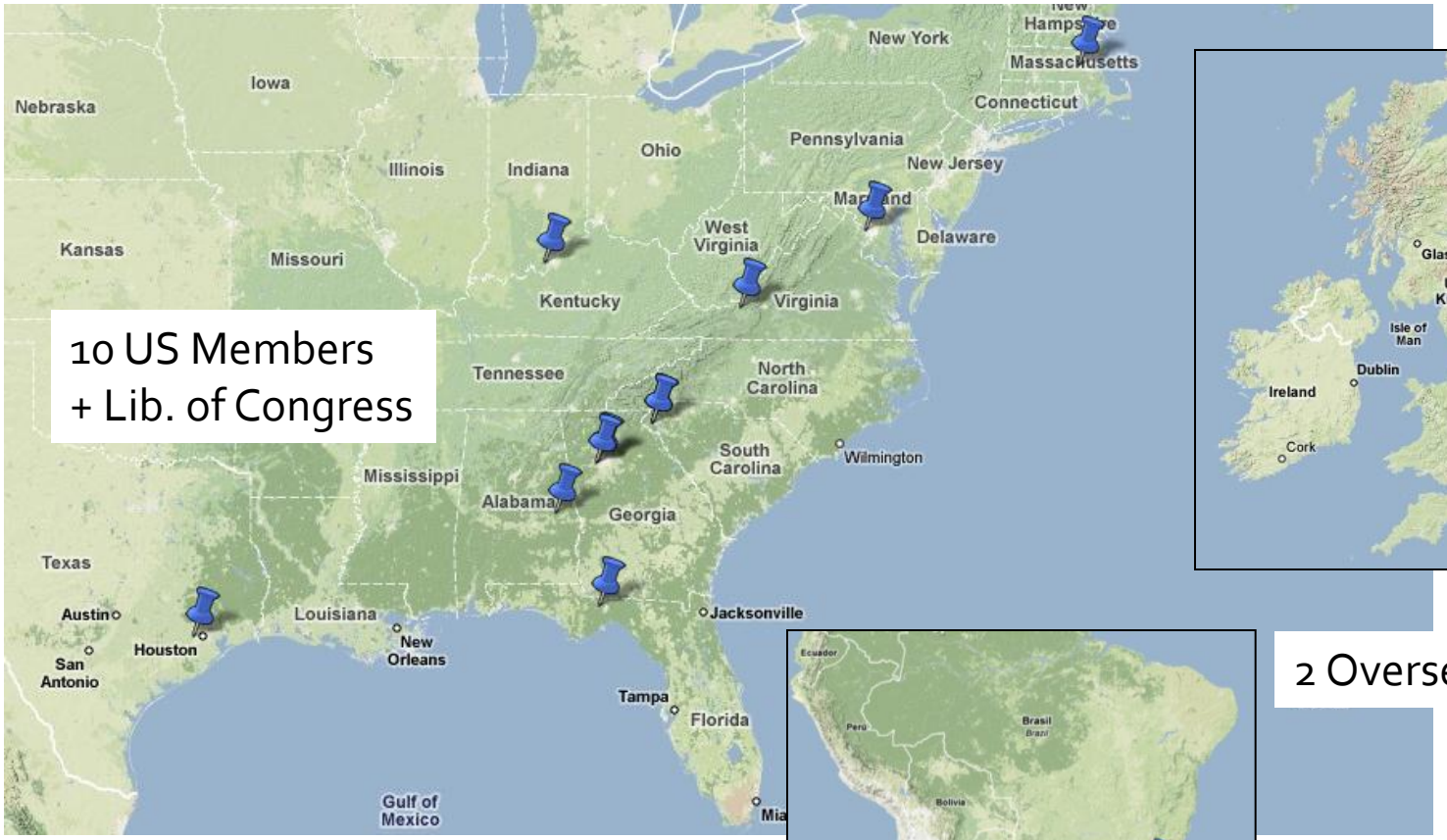
A Community-Based Distributed Digital Preservation Cooperative

- Established in 2003 under the auspices of and with funding from the NDIPP Program of the Library of Congress
- It is both a functioning distributed digital preservation network and nonprofit cooperative for libraries and other cultural memory organizations
- Sustained by cooperative membership fees, NDIIPP contracts, and grants from the National Historical Publications & Records Commission and other groups
- Provides training and models to foster broader awareness of distributed digital preservation and to enable other groups to establish similar networks

MetaArchive COOPERATIVE



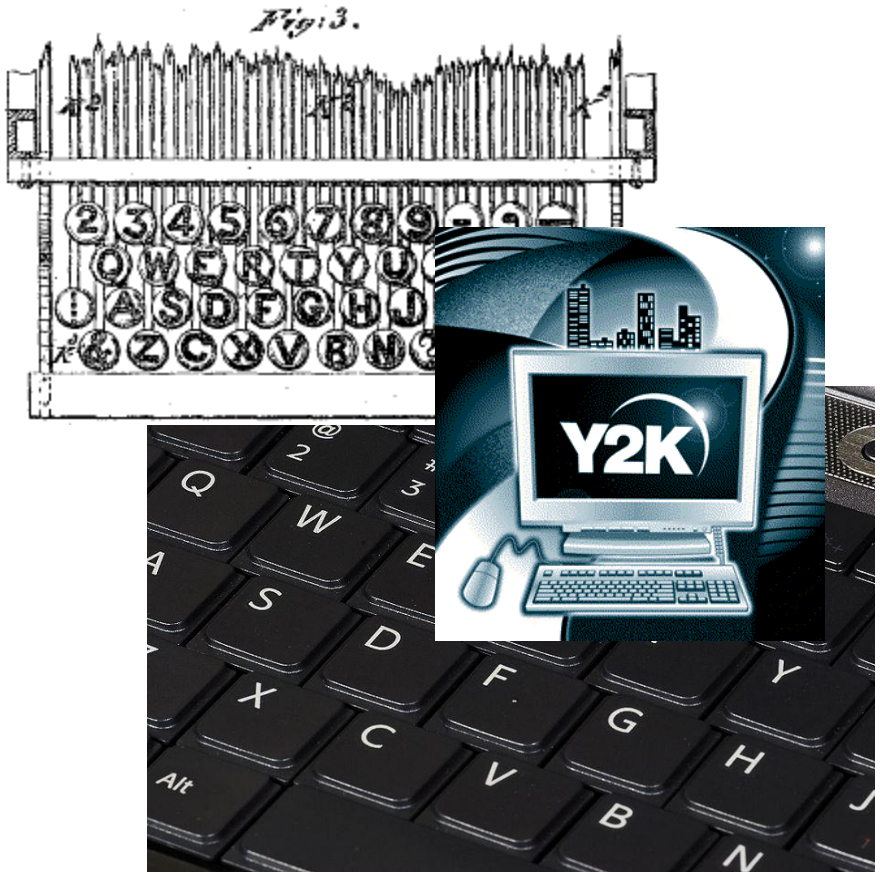
Membership Distribution



2 Overseas Members

Calf-Path Phenomena

- Unfortunate legacies of prior decisions, or lack of decisions/re-examination of processes, are omnipresent and cause enormous problems
- Many (most?) efforts of organizational process remediation are aimed at addressing calf-path issues
- Avoiding following calf-paths should be a goal whenever establishing new precedents



How ETDs get on the Calf-Path

- ETD programs often start with pilot projects featuring ad-hoc procedures and idiosyncratic data storage structures.
- By "data storage structures" we mean the entire range of methods by which ETDs may be stored in structured ways, including directories, administrative metadata, and other data management techniques.
- These idiosyncrasies often quickly evolve into formal practices, much as the awkward and twisted path of the wobbling calf in Foss's poem becomes a standard road followed and solidified by others over centuries.
- Early ad-hoc procedures may become a torturous pathway upon which an academic organization's ETD collection and its management workflows continue to be built.

Recognizing the Calf-Path in ETD Programs

- Do our ETDs accumulate in structures such that we could transfer (preserve) them in logical and discrete groups to another infrastructure, or would such a transfer require restructuring of ETDs?
- Do we accumulate ETDs in patterns that the majority of our staff understands, or do individuals pursue significantly variant processes in silos?
- Are either our ETD storage structures or accumulation processes documented anywhere?

Best Practices: Unique Directory Names

- Standardized, uniform, easy to decipher
- Timestamps
  - etd-mmddyyyy-tttttt
  - http://scholar.lib.vt.edu/theses/available/etd-10022007-144864
  - ETD submitted at 2:48:64 pm on Oct. 2, 2007
- Same naming convention for scanned and born-digital ETDs

Best Practices: File Names

- etd.pdf
  - If file names are not unique, directory names must be unique
  - May not be good for local management
- Lastname_initials_doctype_year.format
  - McMillanGM_T_1981.pdf
  - SoundararajanS_D_2010.pdf
  - SoundararajanS_D_2010_copyright.pdf

Best Practices: Archival Units

- Recommend discreet static (unchanging) archival units (clusters of ETDs)
  - May be done simply with annual ingest into preservation caches
- Suggested accumulation size for archival units of no more than10 GB for portability
  - Divide annual directories into subunits

Best Practices: Triage for ETDs

- Inconsistent practices in directory structures, metadata, and file naming conventions
- Rename, rearrange files or
- Creative strategies needed
- Adapt the existing situation to find, harvest, and ingest the files into the preservation network

Best Practices: Web Accessible

- Keep ETDs on live, spinning discs
- Not on CDs or other static storage devices
- Avoid problems: finding those discs, loading them onto spinning discs, rectifying errors and failed media
  - Even gold CDs regularly fail!
- Declining cost of online storage

Establishing a Digital Preservation Readiness Program

- Start with a shared programmatic vision.
- Document that vision and a corresponding set of best practices for your organization.
- Disseminate your vision and best practices throughout your organization.
- Review your vision and best practices annually.
- Create a registry of collections for your organization.

Recommended Practices for Lifecycle Management of Digital Assets

- Live versus Static Media
- Standardize File and Directory Structures
- Metadata Discipline
- Implement a Digital Preservation Viability and Recovery Program
  - *Assign staff to be responsible for viability and recovery tests.*
  - *Document the entire process of asset recovery.*
  - *Recovery tests should be realistic.*
  - *Conduct periodic tests.*

Contact Info

- Gail McMillan
  - 540-231-9252
  - gailmac@vt.edu
- Martin Halbert
  - 404-727-2204
  - martin.halbert@emory.edu