The Digital Calf-Path:

Growing and Sustaining Digital Collections in the 21st Century

Martin Halbert
(President, MetaArchive Cooperative)

Digital Directions 2009
Wednesday, May 27, 2009
San Diego, California

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 trail was taken up next day
By a lone dog that passed that way;

And then a wise bell-wether sheep
Pursued the trail o'er vale and steep,

And drew the flock behind him, too,
As good bell-wethers always do.

And from that day, o'er hill and glade,
Through those old woods a path was made.

And many men wound in and out,
And dodged, and turned, and bent about;

And uttered words of righteous wrath,
Because 'twas such a crooked path.

But still they followed – do not laugh –
The first migration of that calf.

And through this winding wood-way stalked,
Because he wobbled when he walked.

This forest path became a lane,
That bent, and turned, and turned again.

This crooked lane became a road,
Where many a poor horse with his load,

Toiled on beneath the burning sun,
And traveled some three miles in one.

And thus a century and a half,
They trod the footsteps of that calf.

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.

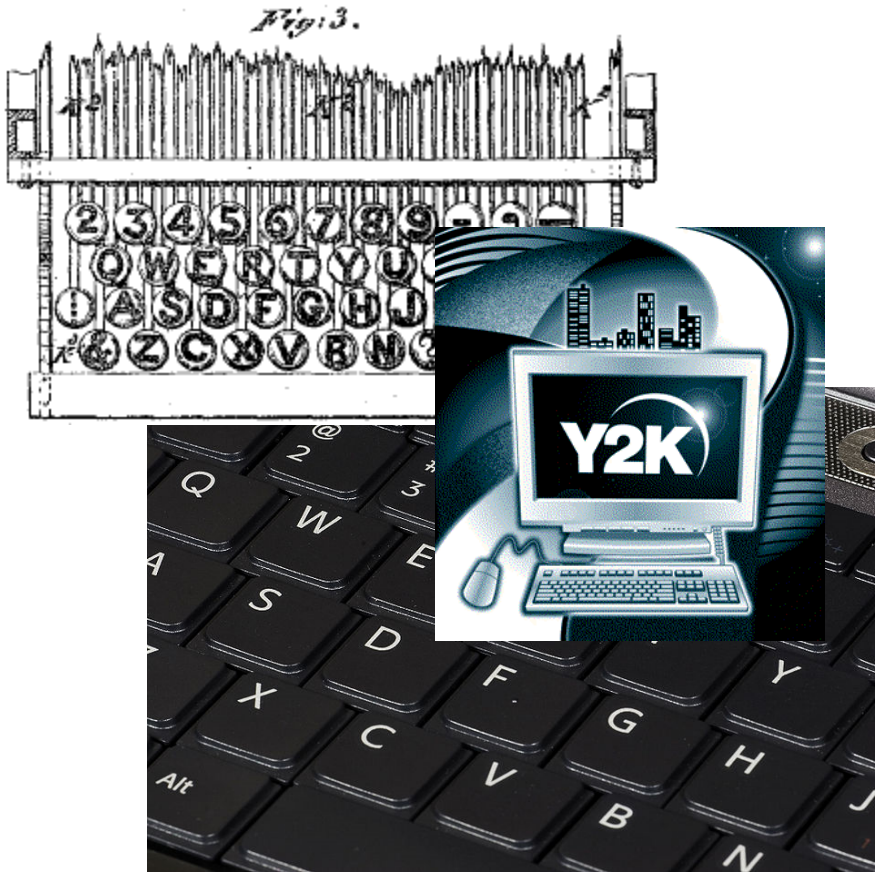
They followed still his crooked way,
And lost one hundred years a day;

For thus such reverence is lent,
To well-established precedent.

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

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



Growing and Sustaining Digital Collections in the 21st Century

1. What are some of the real-world **calf-path problems** that hamstring the long-term survivability of digital collections in "cultural memory organizations" (libraries, archives, museums, and specialized research institutes)?
2. What are the **emerging collaborative responses** of cultural memory organizations to data management problems?
3. What do cultural memory organizations **report about their efforts and needs** to sustain and preserve their digital collections? (survey results)
4. **Recommendations** on specific ways that cultural memory organizations can improve digital preservation readiness

Note: Will focus primarily on digital preservation although most of these comments also pertain to other aspects of sustainability.

Three Calf-Path Case Studies

- The following are some representative actual examples distilled from six years of consulting with cultural memory organization clients that were seeking to address digital preservation issues
- Selected because they are emblematic of particular calf-path problems
- These are examples of bad things happening to good bellwethers who followed practices set down by (someone) before them
- The names have been withheld to protect the embarrassed

Case Study #1: We have a backup program (kindasorta)

- A mid-size cultural memory organization engaged in active digitization of its archival collections for a full decade
- They scanned at high quality and kept reasonable metadata
- They backed up the accumulated scans to tape and stored the tapes…
- …in the room next to the server…
- …and one afternoon ***had a fire and lost it all***.
- ***Observation: local backups are not a preservation program***

Case Study #2:
We know what we've digitized (kindasorta)

- The archives of a major research library engaged in selective digitization of its collections for years…
- …by a series of junior archivists with no official mandate or standardized workflow process in place…
- …the minimal "metadata" captured was recorded idiosyncratically in the image filenames…
- …**after almost a decade they gave up and started over** because no one could now make any sense out of the mass of unorganized files they had accumulated.

Case Study #3: We know where things are (kindasorta)

- The campus electronic thesis and dissertation (ETD) program of a major research university reposited incoming ETD's for years, through a series of workflow processes emulating the original print submission process
- These "echo" processes were created ad hoc by a long series of library paraprofessionals, storing ETD's in one directory, then another, and keeping metadata in non-standardized ***text*** files resembling the paper forms…
- …**after a decade they overhauled the whole program and its workflow** *because they wanted to preserve the data.*

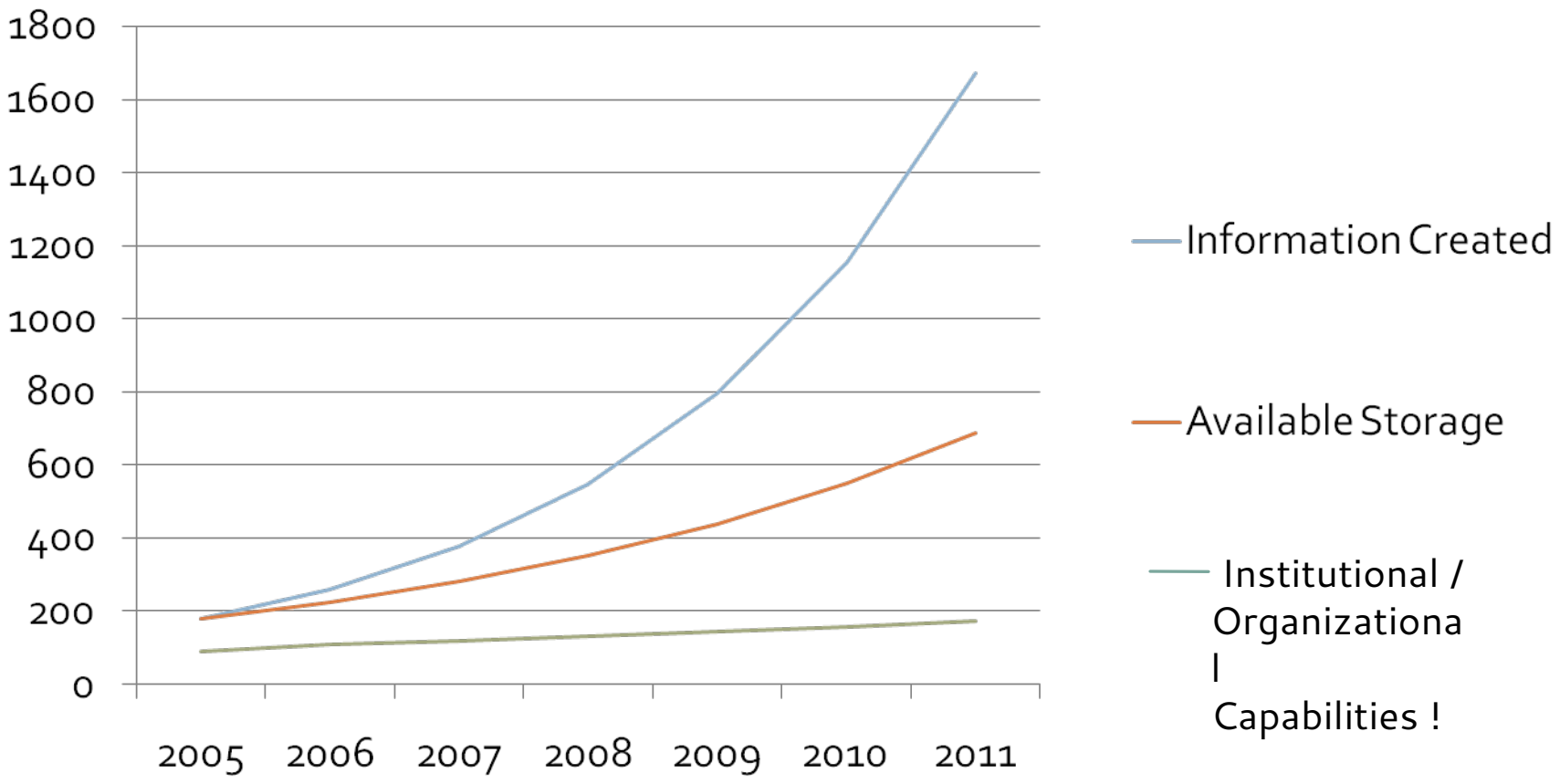
Why Do We Follow the Calf?

- Cultural memory organizations have "deep infrastructure" (professionally internalized practices and organizational systems) with long term management of analog collections
- We lack such deep infrastructure for digital collections because:
  - digital technologies are unfamiliar; they are relatively new and have quite different properties and dynamics
  - digital technologies are unpredictably changing
  - our institutional mandates and funding are often still oriented toward analog, not digital resources
- In an unfamiliar and shifting wilderness, one tends to follow any kind of path encountered, and lack the critical perspective to analytically question such

The Pace of Change Obscures Calf-Paths

Exploding Digital Universe

IDC White Paper 2008



Blue and red chart lines of world data growth in exabytes drawn from ***IDC White Paper*** "The Diverse and Exploding Digital Universe", p. 4

Data Growth will Overwhelm Individual Institutions

- Projections of the growth of data growth are exponential and discouraging
- The organizational and preservation capabilities of our individual institutions are very limited in comparison to this growth
- As individual institutions, we will always be playing catch up with the pace of technological change
- The only way to change this dynamic is to ***change the parameters of the dynamic and act collectively***

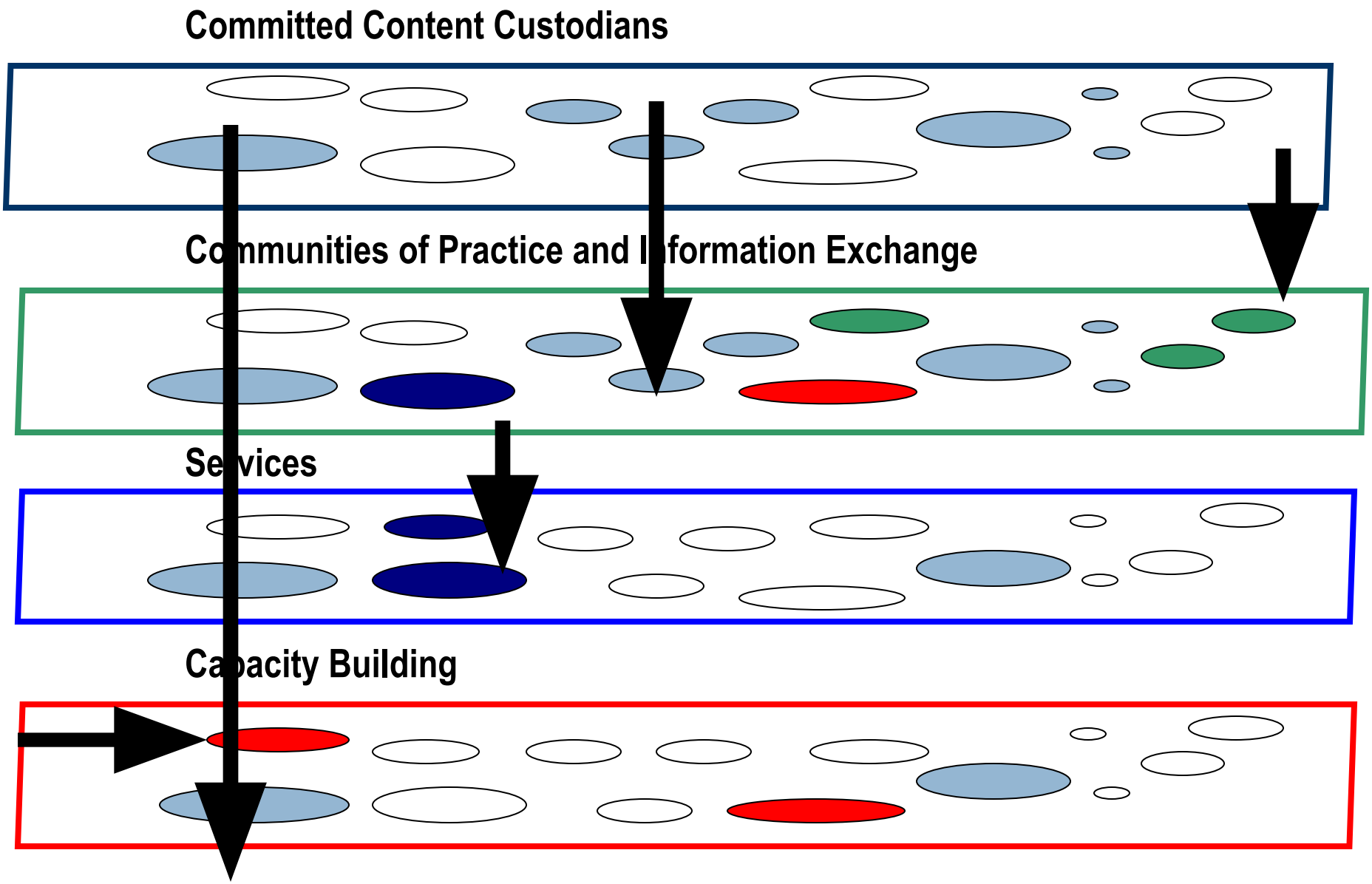
The Necessity of Collaborative Approaches

"The increased number and diversity of those concerned with digital preservation—coupled with the current general scarcity of resources for preservation infrastructure — *suggests that new collaborative relationships that cross institutional and sector boundaries could provide important and promising ways to deal with the data preservation challenge.* These collaborations could potentially help spread the burden of preservation, create economies of scale needed to support it, and mitigate the risks of data loss."

– The Need for Formalized Trust in Digital Repository Collaborative Infrastructure

NSF/JISC Repositories Workshop (April 16, 2007)

Roles in the Stewardship Network



Source: *"Since we met last year..."* Plenary, Martha Anderson, National Digital Information Infrastructure and Preservation Program Annual Partners Meeting 2008

Emerging Collaborative Strategies

1. Building on existing partnerships among cultural memory organizations (often at the state level)
2. Creating new alliances of cultural memory organizations
   - Federal agency initiated
   - Private foundation initiated
   - Independent
3. CMO's align with Profit-making Corporations
   - Cooperative agreements
   - Service contracts and products

Building on Existing State Consortia

- **Attractions** as a strategy:
  - Pre-existing structures, do not need to create something wholly new, and institutions are used to working together
  - Often are already linked to state funding channels for academic and public institutions
- **Questions/drawbacks as a strategy**:
  - May limit the collaboration to the state context
  - State consortia may already be overburdened
- **Examples**: Alabama Digital preservation Network, Pedals program in Arizona

Creating New Alliances: Federal Agency Initiated

- **Attractions**:
  - Funding source external to CMO's
  - Governmental mandate and imprimatur
- **Questions/drawbacks**:
  - What happens when federal funding used up?
- **Examples**:
  - NSF Initiatives (supercomputer centers, DataNet)
  - NDIIPP partners

Creating New Alliances: Private Foundation Initiated

- **Attractions**:
  - Funding source external to CMO's
  - Cachet and prestige
- **Questions/drawbacks**:
  - What happens when foundation funding dries up?
- **Examples**:
  - LOCKSS (original, not current alliance)
  - Technology alliances (Duraspace)

Creating New Alliances: Independent

- **Attractions**:
  - CMO's involved would own the problem
  - Could lead to new professional traditions and practices internalized in CMO's
- **Questions/drawbacks**:
  - Does not bring in funding external to CMO's which are already strapped
  - How to bootstrap this kind of alliance?
- **Examples**:
  - ***None*** (LOCKSS Alliance, MetaArchive, and some other NDIIPP partners come closest but began as federal/foundation projects)
  - Digital Library Federation might have become this if it had continued full bore

Cooperative Agreements with Major Corporations

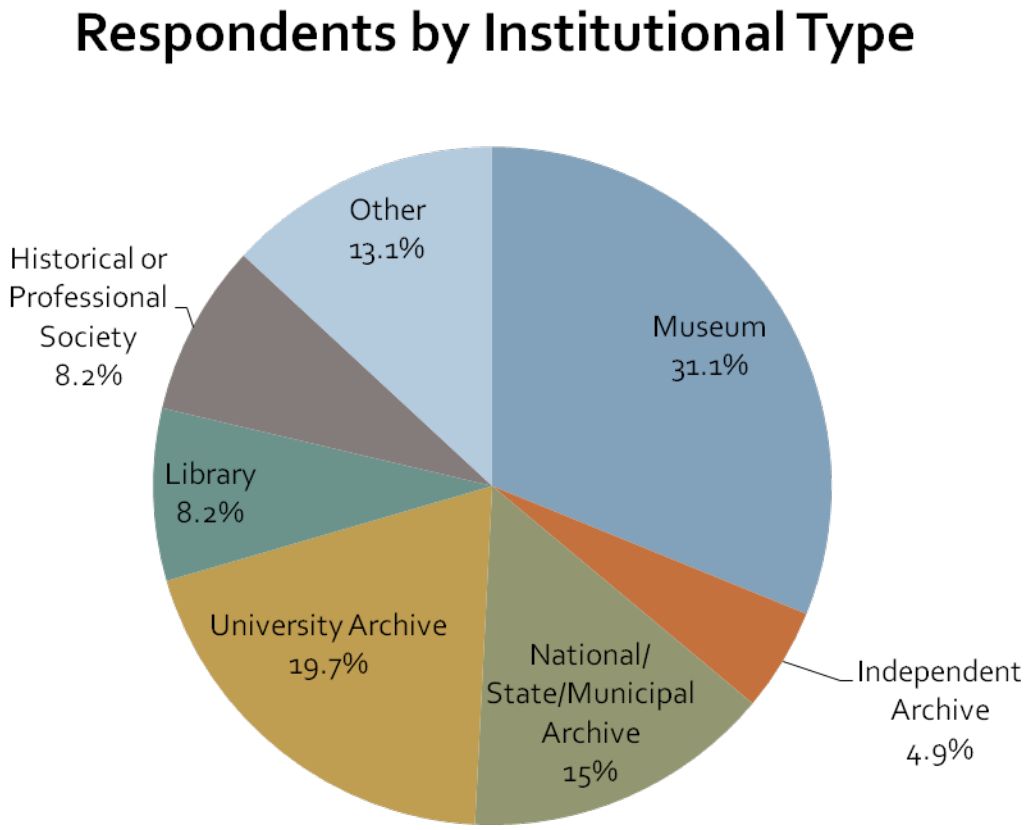
- **Attractions**:
  - Enormous capitalization of corporate partners can bring in major external funding
  - Tremendous technical expertise and infrastructure
- **Questions/drawbacks**:
  - Imbalance of power in the agreement
  - Historical example of the journal crisis that arose when corporations gain monopolistic control of information
- **Examples**:
  - Google Books project and Hathi Trust
  - Microsoft scanning project

Service Contracts with Specialized Vendors

- **Attractions**:
  - Longstanding contractual relationships / contracts with library vendors is reassuring
  - Outsourcing is perceived as an attractive option because it has the aura of "optional expense that can be cut" to administrations
  - Attractive for CMO's that lack technical expertise in-house
- **Questions/drawbacks**:
  - Is preservation of digital collections something that should be outsourced to vendors? Would we outsource preservation of our analog archives?
- **Examples**:
  - Portico for e-journals

What do Cultural Memory Organizations Want? What are they doing?

- The MetaArchive Cooperative conducted a survey of CMO's in 2009 Q1 of ***Digital Preservation Practices and Priorities***
- The purpose of the survey was to assess the digital preservation desires, preferences, and needs of CMO's
- 60 responses



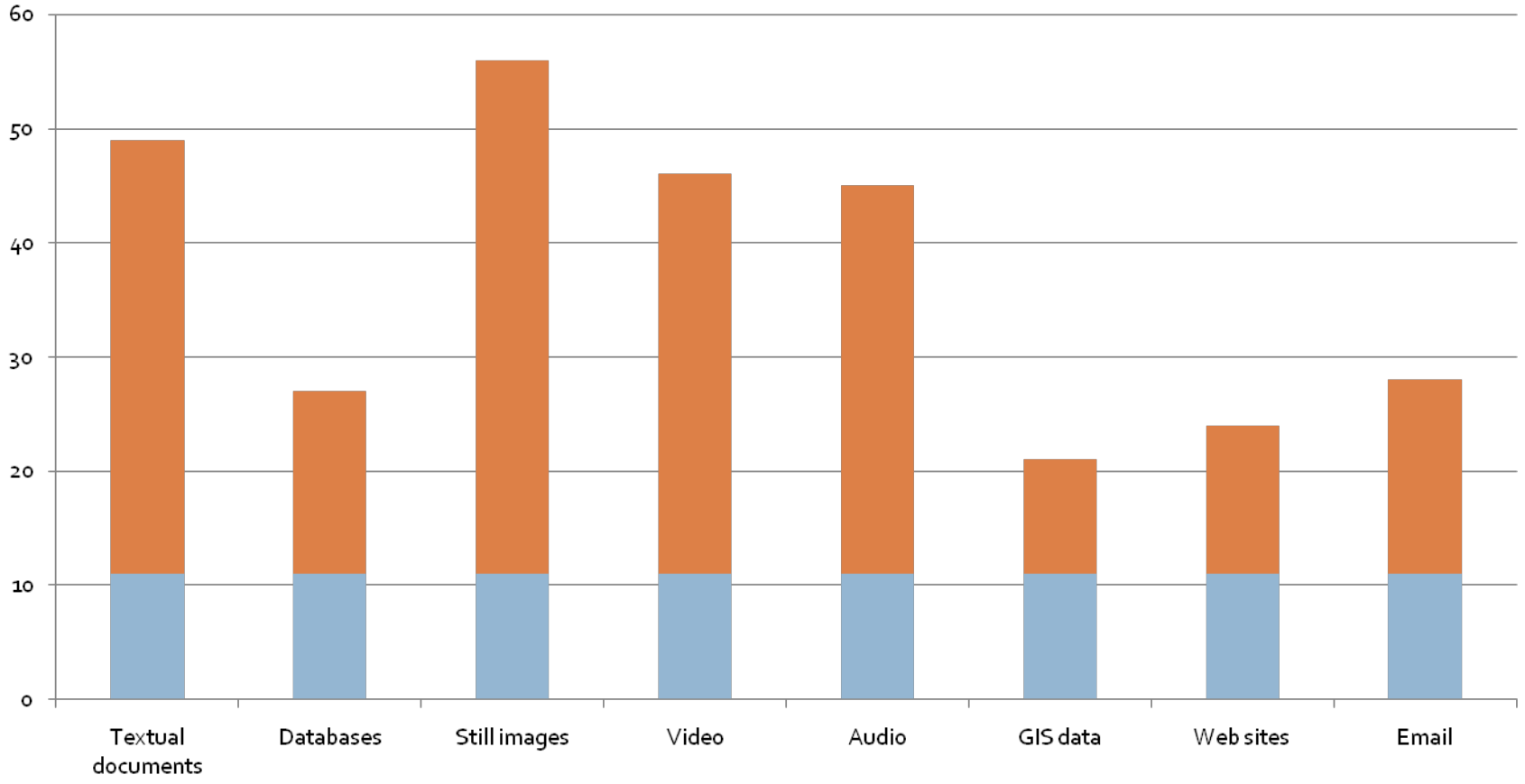
Growth of Digital Collections

- **Institutions are now actively acquiring and building significant digital collections, even at small museums and archives**
- Fully 94.8% of respondents indicate that their archives has a growing digital component.
- The average size is 2 TB, with 24 respondents reporting 1 TB or less and 22 reporting between 1 and 20 TB
- Collections are quickly growing at an average of 540 GB/year
- Twenty-six respondents report that they expect to add 500 GB or less and 19 anticipate adding more than 500GB in the next year

Diverse Formats

- **These institutions need to preserve a diverse array of format and genres.**
- Still images are cited by respondents as the dominant format (94%) . Other popular formats include textual documents (83%), video (76.2%), and audio (74.5%).
- A significant number of respondents also collect email (47.4%), databases (47.5%), and websites (40.6%), and more than a third also collect GIS material (35.5%). Others cited presentation materials, publications, science data, and software code as items that they collect/create.
- **Presumably, as these institutions look for preservation solution(s), they will be seeking solutions that can encompass multiple formats and genres of material easily.**

Formats Distribution



Wide Variation in Infrastructures

- **We found that the platforms and repository structures used by respondents varied widely**
- More than half (65%) of the respondents are using an in-house solution to host some or all of their collections.
- The leading repository systems cited were CONTENTdm (9 users,17%), Fedora (5 users, 9%), DSpace (4 users, 7%), and Access/Excel (3 users, 6%).
- Two each cited SRB and Filemaker as repositories that they currently use. Ten other systems were cited by ten additional respondents.
- **This diversity of system types—presumably at least somewhat representative of the overall industry—presents an array of challenges for preservation, especially for distributed replication approaches.**

Inadequate Policies

- **Institutions need assistance in designing suitable policies for managing their digital materials.**
- *Only 21% of respondents state that their institution has a preservation plan for its digital archives.*
- Less than half of respondents have any written policies for managing their digital collections (44.8%).
- ***Of those that have policies at all***, most only cover metadata standards (76.9% %), back-up strategies (65.4%), conversion of materials from print to digital (61.5%), preservation (57.7%), and acquisition (53.8%).
- ***Only three institutions (5% of respondents) believed that they had policies that met their***

- **Many of these institutions are not yet even backing-up—let alone preserving—their digital assets**
- Only 25 respondents (50% of 50 respondents) stated that they back up all of their digital holdings
- Of the remaining, less than a quarter (11) back up 75% or more, only 5 more back-up 50% or more, and 8 back up less than 50%.
- Only 10 respondents believed they had in-house expert knowledge regarding digital preservation, 36 felt they had intermediate knowledge, and 8 believed they had only novice-level knowledge of

Biggest Threats

- **Institutions perceive the insufficiencies of their policies, plans, and resources to be the biggest threats to the loss of their digital assets**.
- According to survey responses, the greatest threat to collections is insufficient resources for preservation (75%, or 41 institutions), followed by insufficient policies and plans for preservation (51%, or 28 institutions).
- Less troubling to these institutions were such issues as technological obsolescence (31%, or 17 institutions) and the stability of the storage medium (20%, or 11 institutions).
- Institutions need affordable, prescriptive options.

Interest in Cooperative Preservation Strategies

- **There is widespread interest among cultural memory organizations in participating in a cooperative preservation network.**
- Three-quarters of respondents (42 institutions) cited interest in participating in a cooperative preservation network, and 89.5% cited interest in "participating in a community-based digital preservation solution." This stood in contrast to those who indicated interest in preservation services provided by third-party vendors (30.4%, or 17 institutions).
- There is also widespread knowledge of and experience with LOCKSS among the respondents, with almost 50% (28 institutions) citing such knowledge and experience.
- This indicates that the community widely supports the idea of maintaining an active role in collaborative digital preservation activities.

Typical Elements of Collaborative Approaches to Digital Preservation

- Need to identify:
  - Digital preservation approach (centralized, distributed, formats, genres, etc.)
  - Organizational framework and governance (based on existing or new consortia)
  - Roles and responsibilities
  - Technical standards
  - Funding model

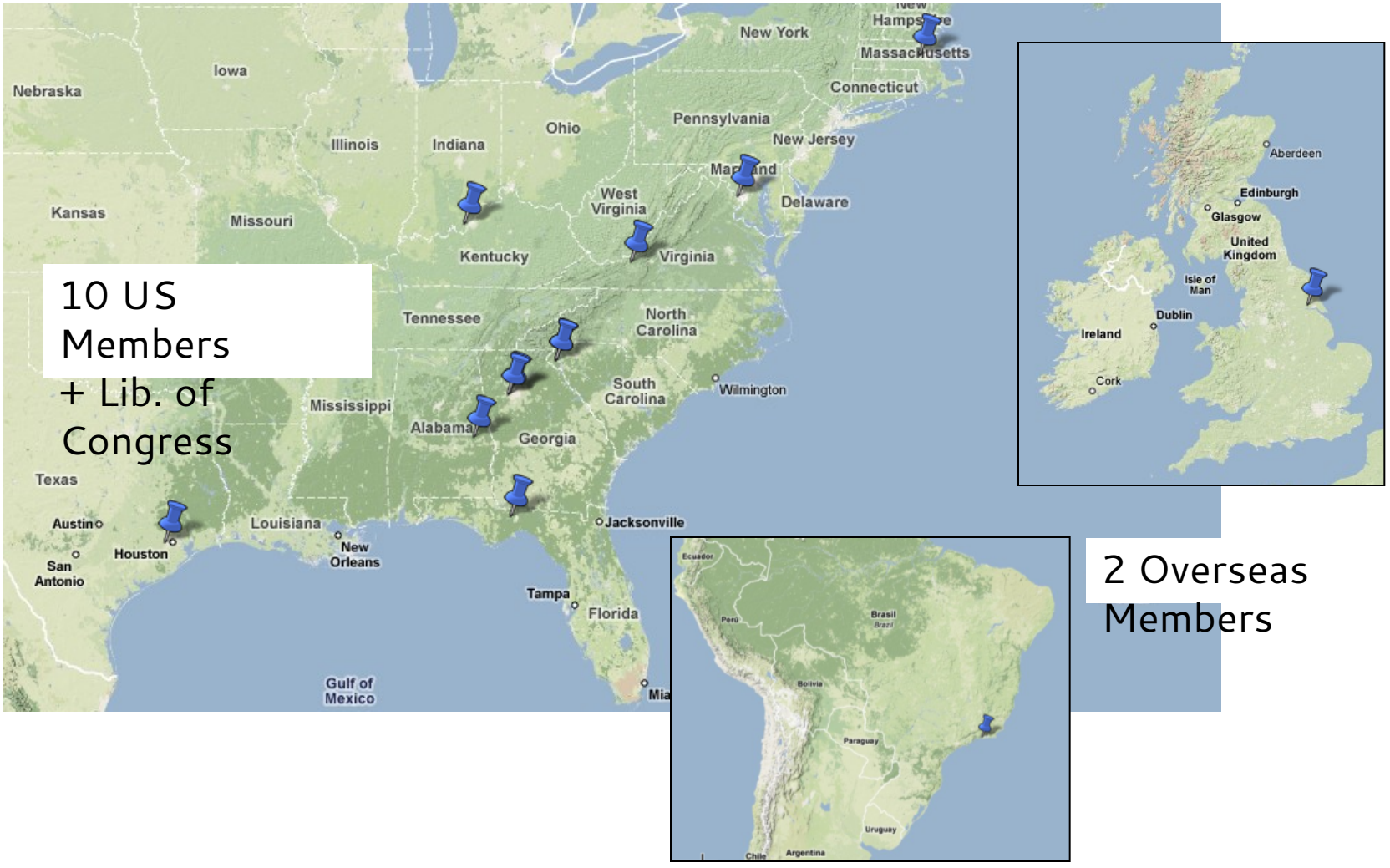
MetaArchive Case Study: A Community-Based Distributed Digital Preservation Cooperative

- The MetaArchive Cooperative was established in 2003 under the auspices of and with funding from the National Digital Information and Infrastructure Preservation 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

Preservation Approach: a Distributed Shared Infrastructure

- MetaArchive provides a cost-effective option for distributed digital preservation through a cooperative structure:
  - secure distributed digital preservation of archives
  - based on re-use of LOCKSS for digital archives
- Participation as a MetaArchive node is relatively simple and cheap
- Benefits of membership are understandable by institutions, bounded, affordable, and address an important institutional gap

Membership Distribution



10 US Members + Lib. of Congress

2 Overseas Members

Current Members

- Auburn University
- Boston College
- Clemson University
- Emory University
- Florida State University
- Folger Shakespeare Library
- Georgia Tech
- Library of Congress (Sponsor)
- Pontifical Catholic University of Rio de Janeiro
- Rice University
- Hull University Wilberforce Institute
- University of Louisville
- Virginia Tech

Technology: Building on Top of LOCKSS as a Solution for Preserving Digital Archives

- Conspectus Database (Original)
  - Curators enter collection level entries for collections
  - Meant to be used for cooperative prioritization in DDP selection and decision-making activities
  - Not interactive with some key MetaArchive systems (Cache Manager, Ingest Plugins)
- Second Generation Conspectus Database
  - Now in development
  - Integrates operation of all network functions
  - Being designed in concert with guidance from other PLNs, hopefully in ways that enable re-use

Organizational Agreements and Models

- Developed a new cooperative with guidance from both legal team, librarians, and intellectual property specialists
- Created core organizational documents in 2006: charter, membership agreement, papers of incorporation, business plans, etc.
- Allows members to understand their commitment and liability clearly

Collection Foci: Growing Archives in Subject & Genre Domains

- **Southern Digital Culture** (initial collecting area, founding members were Southeastern)
- **Transatlantic Slave Trade Historical Data** (made cooperative international)
- **Electronic Theses and Dissertations** (inter-consortia strategic alliance with NDLTD)
- **Early Modern Literature** (broad new area, with Folger Shakespeare Library as cornerstone)

Active Collaborations with Other Efforts

- **LOCKSS** (collaborative development of LOCKSS Cache Manager)
- **Data-PASS Alliance** (developing in-common standard for Private LOCKSS network interoperation standard and tools)
- **ECHO DEPository Project** (PLN interoperation standard using HandS)
- **SDSC Chronopolis** (PLN/ SRB interoperation testing and bridges)

The Inception of a New Field

- We need a healthy mix of models and options for successful digital preservation
- The field is still emerging, but now is the time to think constructively and analytically about how we want it to develop
- Organizations now forming must make strategic decisions about how they want to either synergize or compete with

Conclusions: Recognizing the Calf-path Syndrome

- Digital objects and metadata are embedded in a closed system from which they cannot be effectively extracted in a coordinated way
- Various digitization streams are structured by ad hoc decisions of staff with unpredictable patterns
- There are limited or no metadata other than file naming conventions and staff memory of what file names mean

Diagnosing the Calf-Path

1. Do our data assets accumulate in structures such that we could package them up and transfer them to another infrastructure in a straightforward way, or would such a transfer require ad hoc bundling?
2. Do we accumulate data assets in patterns that the majority of our staff understands, or do individuals pursue significantly different processes in silos?
3. Are either our data storage structures or accumulation processes documented anywhere?

Recommendation: Establish 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.

Finally: 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.*

The Way Forward

- There is audacity in hope for the future – ***there is strength in this audacity***
- There is more that draws us together than separates us
- We as a community of cultural memory institutions share many values concerning our commitment to preserving a public web of knowledge, values that motivate us to work together cooperatively

Contact Info

- Dr. Martin Halbert
- 404-727-2204
- martin.halbert@emory.edu