Avoiding the Calf-Path: Digital Preservation Readiness for Growing Collections and Distributed Preservation Networks

Martin Halbert, Katherine Skinner; Emory University; Atlanta, GA. Gail McMillan; Virginia Polytechnic Institute and State University; Blacksburg, VA.

Abstract

Over the past six years, the members of the MetaArchive Cooperative have worked to identify a series of best practices for distributed digital preservation readiness. These best practices can benefit ongoing initiatives as well as start-up programs which have not yet established regular procedures and standards for directory structures, metadata, and file naming conventions. We document what we term the "calf-path syndrome," the way in which early strides in an organization's digitization work may create a legacy that is detrimental to the preservation readiness of their growing digital collections. We share relatively simple principles and guidelines for such programs that can greatly improve the subsequent likelihood of implementing successful distributed digital preservation programs.

Introduction

A hundred thousand men were led
By one calf near three centuries dead.
They follow still his crooked way,
And lose one hundred years a day,
For thus such reverence is lent
To well-established precedent.
-Sam Walter Foss, "The Calf-Path"

Libraries and other cultural memory organizations regularly create major digital collections as part of their ongoing work. The genesis of these collections is often a series of iterative and cumulative digitization efforts with idiosyncratic and ad-hoc data storage structures. By "data storage structures" we here mean the entire range of methods by which data is stored in structured ways, including directories, administrative metadata, and other data management techniques. Much like the awkward and twisted path of a wobbling calf that becomes a standard route followed and solidified by others over centuries, the early idiosyncrasies embedded in these collections' data structures can become a torturous pathway upon which an organization's digital infrastructure and its management workflows continue to be built.

Such infrastructures may cause curators enormous problems when they engage in systematic efforts to digitally preserve the content of growing collections. We address these problems by providing practical suggestions, recommendations, and guidelines for institutions that find themselves on a "Calf-Path" of their own making. These recommendations are informed by six years of practical experience in addressing such issues in the course of operating the MetaArchive Cooperative, a distributed digital preservation cooperative of cultural memory organizations. While there are many potential strategies for preserving data over long periods, we fundamentally believe that all effective strategies will include some kind of secure and distributed replication of the data in question, and our discussion will focus on readiness for such distributed digital preservation activities. We also differentiate repository programs from digital preservation, as we consider repository systems a means of managing workflow and access to digital collections rather than a full digital preservation strategy.

Why Institutions Follow the Wobbling Calf

The early part of the 21st Century marks an unprecedented historical moment in which cultural memory organizations are increasingly expected to preserve a variety of digital collections, ranging from the output of analog-media digitization projects to the administration of born digital collections such as electronic theses and dissertations (ETDs). Most cultural memory organizations (like many other institutions in society) lack deep experience with digital infrastructures (what Don Waters and others have called "Deep Infrastructure" for short [1]). Many digital repository efforts have humble beginnings with limited resources and staffing, but may gradually grow over time to produce and process very significant bodies of content and volumes of material.

Two categories of repositories that have experienced such growth and which we will here consider are a) programs that engage in progressive digitization of print archives, often funded sporadically through grants, and b) ETD repository programs, which often begin with mostly unfunded institutional mandates in universities. We will discuss a variety of findings concerning digitization programs and examine ETDs in more detail through a case study from MetaArchive. Both of these content categories tend to grow in an effectively unbounded manner over time. Sporadic archival digitization projects often lead cumulatively to collections created using irregular and varying practices determined by ad hoc exigencies of individual projects. ETD repository programs are also often driven by exigencies associated with creating an effective electronic workflow for accepting and securely storing digital copies of theses and dissertations as either a replacement or supplement to parallel workflows for print copies. As these repositories seek to preserve their digital collections, they may find that their collections' directory structures, naming conventions, and other structural organization elements limit the preservation readiness of these collections.

For example, an organization might begin creating a digital archive through a sponsored-funding supported newspaper digitization project. As part of the project, the organization creates a directory structure and naming conventions that make sense within the narrow confines of the newspaper content base. A year later, different individuals within the same organization might gain funding to digitize and encode a collection of rare books in XML. Radically different directory structures, naming conventions, and organizational practices might be implemented for this new digital

content. Why not? This is still a green field, and we all wobble when we first learn to walk. In the absence of an agreed consensus on standards of practice that comes with deep infrastructure, each new digitization effort lays down a pioneer's trail that subsequent projects may or may not follow. Over time, the organization may create and acquire many more digital collections through ad hoc projects, each time developing new organization conventions, or worse, creating undocumented variations on those originally set forth in that first project.

In this way, multiple "calf-paths" are created in cultural memory organizations, especially by groups with different professional practices, one by archivists, another by digital librarians, and yet another created by records managers. They all belong to the same organization and share a need to preserve the various collections over time, but the task may have become essentially impossible because of the calf-paths that the various groups have become habituated to using.

Eventually, the organization accumulates a sufficiently large and valuable set of collections that a) the variations in organizational practice become untenable to sustain, and b) the content stewards realize that they need to actively curate and preserve these digital collections. A review of the accumulated variations in collections' directory structures, naming conventions, and metadata forms finds that they are idiosyncratic, outmoded, and hindering the preservation readiness of the organization's digital assets. Yet, these organizational practices are like a long-established twisting maze of streets: daunting to consider overhauling. Remediation would cost almost as much as redoing the original digitization projects.

It is important to understand that in observing this phenomenon, we are neither casting aspersions nor proclaiming our own virtue. We have observed the condition that we here term the calf-path syndrome in virtually all the institutions of MetaArchive to some degree or other, as well as virtually every other cultural memory organization engaged in digitization with which we are familiar. This is not aberrance. It is the normal state that virtually all of us in cultural memory organizations find ourselves in during the still early decades of the digital age. The question is: *what do we do about it?*

Recognizing the Calf-Path

The most important step in addressing a problem is to diagnose its existence. As stated above, almost all cultural memory organizations are experiencing a calf-path syndrome to some extent at any given moment. While we all may note aspects of this problem in passing, any specific data organization snarl is typically not critical enough in comparison to day-to-day exigencies and deadlines to prompt the kind of overhaul that would address the overall accumulation of problems. There is rarely a trigger event grave enough to make an organization set aside immediate priorities for long-term benefits.

The first step in responding to of the calf-path syndrome is becoming cognizant of its existence through some kind of self-assessment. We therefore offer the following set of questions concerning digital preservation readiness, which we suggest organizations should seriously consider asking periodically about their digitization programs:

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?

We are here focused on digital preservation readiness because we have found that often the first time that we acknowledge the long-term detrimental effects of the calf-path syndrome is when we seek to preserve our digital assets. Digital preservation, after all, is not simply the process of keeping bytes of content technically alive and viable. As important, those bytes of content must still be understandable and renderable; they must make sense to human eyes. And it is of no help if an organization "preserves" all of its collections without structuring them such that they can actually be used to repopulate that organization's infrastructure in the event of data loss.

For example, consider an institution that has a series of collections, each stored in esoteric formats with different naming conventions, irregular directory structures, and metadata in various undocumented schemas. The institution exports those collections in their present forms to a distributed digital preservation network. Those collections will be "preserved" in the sense that all of the bytes will be retrievable if that organization should need them in the future. But if calamity strikes and that organization indeed retrieves its collections from preserved storage, how will they know where the files belong? How will they recreate their diverse collections if they have not carefully documented their structures prior to preservation? And if there is no working import process mirroring the export process, how will the functioning archive be recreated? Imagine the curator looking at an item "15326b.jpg" stored in a collection of 960,000 objects that include all of the "digital masters" created by that organization, with a separate dump of descriptive metadata not mapped to the filenames of the masters. Is the object technically preserved and viable? Perhaps. Is it useful in its current state, divorced from the context of its creation? Absolutely not. If we are honest with ourselves, we will realize that we have followed the calf-path into the deep weeds.

Once we recognize that we are following a set of precedents that are not supportive of our stewardship goals, we have the opportunity to move beyond a twisting path to establish something that closer resembles a roadway, as will be explained later.

All too often, institutions simply dismiss the calf-path syndrome as an unavoidable and unquestioned received legacy of the period in which the digital collections were first created. But we reiterate: *the first step is seeing that a calf-path is there*. Common symptoms by which the calf-path syndrome can be recognized include:

- 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

We have seen variations on all of the above problems in archives of print digitization projects, and each problem presents obvious difficulties when attempting to preserve content in

meaningful ways. Some of these problems are intractable, and very significant remediation is necessarily required, ranging from basic cataloging and processing of images to wholesale reorganization of archives.

Sometimes, however, there are relatively straightforward and economical ways of remediating the calf-path syndrome. We will consider a case study of electronic theses and dissertations (ETDs) as a practical example of a type of organizational content that grows in unpredictable ways subject to the calf-path syndrome, and which is in great need of preservation efforts.

Lessons Learned: MetaArchive Experience with ETD Distributed Preservation

The MetaArchive Cooperative and the Networked Digital Library of Theses and Dissertations (NDLTD) formed a collaborative alliance in 2008, in part because both organizations believe in helping higher education institutions provide long-term open access to ETDs. To determine that there was a need for and an interest in a distributed preservation network for ETDs, we invited participation a survey in 2008 through listservs aimed at library and graduate school leaders, including the Association of Research Libraries (ARL), Association of South-Eastern Research Libraries (ASERL), Council of Graduate Schools (CGS), the Digital Library Federation (DLF), and NDLTD.

In three months, this 18-question multiple choice and short-answer survey garnered 95 responses. It revealed that 73% of the institutions do not have formal preservation plans for their ETDs. Nearly all were interested in "participating in an ETD-specific LOCKSS-based collaborative distributed digital archive sponsored by the NDLTD." In fact, nearly three-quarters indicated that they would want to "share preservation responsibilities by running a secure server for the network" and one-quarter would want to have an active role in the MetaArchive by contributing "to the growth and maintenance of this network both technically and organizationally." The MetaArchive Cooperative and NDLTD have therefore begun a pilot project to examine the practical issues involved in a collaborative replication strategy for digital preservation of ETDs. Our findings have been illuminating, and are worth reviewing to understand the syndrome.

ETDs on the Calf Path

Whether an ETD initiative is well underway with required submissions or has just begun with a few voluntary submissions, long-term preservation is among its goals even when a specific plan is lacking. Backup file systems are usually in place for ETDs and other digital resources, but an actual strategy to provide long-term access to these records of university research is often left on the back burner while issues such as workflow from submission to approval and storage are established.

Directory structures and file naming conventions of ETD collections are frequently created without considering their impact on the collections' preservation readiness. All too often, primary consideration is first given to local storage issues and near-term access. As a result, ETD collections often grow almost arbitrarily, seemingly structured but lacking the logic or hierarchy that favors subsequent distributed preservation and access strategies. For example, we have found that ETDs are often simply stored in one mass upload directory, rather than being structured as files in manageable clusters, such as yearly accumulations. When it comes time to preserve such collections, it is difficult to establish what of the collection should be preserved. Questions quickly emerge. How can the institution create a Submission Ingest Package (SIP) for a moving target, one that continues to grow within the same file folder in an unstructured manner? When should they prepare their next SIP, and what should it contain? If it contains a full replication of the folder, it means that the same files are being preserved in multiple instances (begging the question someday of which is the master file), but if it does *not* contain a full replication, how can the institution be sure that it has captured everything that it has added to the folder? Add to this question the details about how to handle embargoed files, files that have been removed or changed, and how to account for retroactively scanned theses and dissertations as well as new born-digital works, and the thicket of the institution's calf-path becomes evident.

Experience in the NDLTD/MetaArchive pilot project has enabled us to make the following suggestions for best practices for ETDs. These need not be implemented only by institutions just starting their ETD collections; they may also be adopted as a more straightforward path for institutions already in mid-stride. In these cases, if there is not yet time for remediating older files, at least the files created in the future will be geared toward long-term access and preservation readiness.

ETDs: Recommendations and Best Practices

Effectively organizing an ETD collection for preservation requires creating a broad-based logical structure such as a directory for each year's ETDs. We recommend that large institutions that add hundreds of files annually subdivide their annual directories into further logical units such as semesters or months. Adopting a uniform, regular, and easy to decipher naming convention for files is also helpful, for example year/month would be 2008-01, not 2008-January, then 2008-Feb, etc.

Any effective digital preservation strategy must impose some practice for automated and therefore structured wrangling of content into manageable packages (SIPs). In the jargon of the LOCKSS software which MetaArchive uses for secure distributed preservation, such packages are referred to as Archival Units and are conceptually fundamental to a systematic replication process designed to comprehensively preserve all the ETDs in question. Whereas structures optimized for human browsing might be based on departments, authors, advisors, etc., an organizational approach designed for comprehensible workflow and preservation of a growing collection is more usefully based on accumulation periodicity.

Triage for Legacy Collections

Making a clean start with orderly structures and practices is the best option, but what about collections that have followed the calf-path for years? Such older collections may require creative strategies. Triage may call for data wrangling to mitigate cumbersome collections and rearrange files into a predictable order so that the ingestion path can be clearly defined. It is less-than-desirable to move and rearrange files, but this can lead to discovering missing, mis-numbered, duplicated, etc. files. Identifying and correcting these problems will, of course, help not only with preservation, but also to improve local access.

When it is impractical or an institution is unable or unwilling to move and/or rearrange files, it is still possible to adapt the

existing situation to find, harvest, and ingest the files into the preservation network. The first adaptation is to cease adding to this collection, thus creating a static collection with a now finite number of cumbersome files, and begin to implement new best practices based on the above logic.

Case Study: Early Virginia Tech ETDs

A small case study is instructive here. Virginia Tech's 1996-1999 ETDs are an example of the calf-path syndrome. These ETDs were reposited using a variety of URN conventions, such as /etd-454016449701231/ and /etd-030999-145545/ Students who submitted their ETDs through the ETD_db prior to the major software upgrade in 2000, were still, however, assigned unique identifiers--URNs, but they were not consistently structured from today's point of view and went through several iterations. For example, a dissertation labeled 030999-145545 was approved in 1999 and one labeled etd-454016449701231 was approved in 1997. As a means of remediating this heterogeneous collection, Virginia Tech created a *virtual and artificial* collection with just one archival unit for all pre-2000 ETDs. The plugin software for the archival unit is instructed to find all ETDs that do not fit the post-1999 URN convention. The complexity of this static collection is best served by plugin rules that exclude anything that matches the post-1999 format structure and places it into an "Early VT ETD" Collection.

The recommended Virginia Tech naming convention for ETDs now follows the format *etd-mmddyyyy-ttttt* and is based on the timestamp when they are added to the collection. Some ETDs from the calf-path era may include unpadded months and days as well as two- and four-digit years but these are also correctly harvested into AUs by year. Therefore, anything that does not match this file naming convention can become a separate collection. At Virginia Tech this collection is simply named ETDs@VT - pre 2000 unsorted and we use the plugin name edu.vt.library.thesesearly. This collection is static and no new ETDs are added with the inconsistent file naming conventions. This collection also harvests the non-ETD content in the /theses/ directory because it excludes pages that follow the above format.

Scanned (versus born-digital) theses and dissertations follow the recommended file naming convention based upon their digitization date, not the original date on which they were approved. This allows the static collection to remain unchanged. This system works for preservation purposes; however, it needs further consideration for rebuilding a public ETD database or collection from the preservation cache because works cannot be programmatically identified to reestablish an annual grouping based on year of completion/approval.

This example illustrates a general strategy of remediation: recognizing and putting boundaries around an irregular collection as a calf-path area that requires special measures for data management. Short of reprocessing the entire collection retroactively (the equivalent metaphorically of bulldozing the calf-path and starting over), a reasonable strategy is to isolate it with special signage and create a roadway going forward.

From Calf-Path to Roadway

So, assuming an organization has recognized a calf-path in its midst, what should it do going forward? We believe that an initial helpful activity is to develop a digital preservation readiness program. To reiterate our starting perspective, we feel that the most effective preservation strategies incorporate pre-coordinated replication of content in distributed and secure locations. As we have discussed, in the digital realm, such replication strategies become increasingly difficult to implement when the content is stored using irregular practices in directory structures, metadata, and file naming conventions; in short, when digitization efforts are trapped on a calf-path. As cultural memory organizations seek to engage distributed preservation strategies, what becomes apparent is the need for clear guidelines to help them structure collections for preservation readiness.

As we have brought new organizations into the MetaArchive Cooperative, we have carefully considered such guidelines as a way of building organizational readiness for distributed preservation activities. During the last six years we worked with our member institutions to articulate principles and guidelines for such programs that can greatly improve their preservation readiness. We feel that these best practices can benefit start-up programs and also help established programs to restructure.

Establishing a Digital Preservation Readiness Program

Designing a digital preservation readiness program that incorporates information about standard means of collecting and storing files is fundamental to an institution's preservation readiness. But how do you establish a program? And how do you ensure after its creation that it will not become a dusty, misnamed set of files buried in a directory tree under which no staff member has any hopes of finding it?

The Cooperative and its members have designed a five-step process to preservation planning on the basis of their work. The steps are as follows:

1. **Start with a shared programmatic vision.** The key word here is "shared." From assessment to publication, the Program should be designed by representatives from across the organization. It should also have the buy-in of the organization's head, be that a director or a board of directors.
2. **Document that vision and a corresponding set of best practices for your organization.** The documentation you create should be easily accessible to members of the organization who are already involved in digital preservation, and also to those who become involved in the future. Document the collections that need to be remediated to fit this new set of best practices, even if there is no funding now to begin this work.
3. **Disseminate your vision and best practices throughout your organization.** Do not ascribe to the "build it and they will come" model. Ensure that the documentation is well known by your staff through presentations, newsletter blurbs, and wide dissemination.
4. **Review your vision and best practices annually.** Keep the documentation alive through dating its production and scheduling annual reviews by the organization. All of us know that the digital landscape is ever changing in this early phase of its development. Your processes should be flexible enough to change when needed, but those changes should be checked in and documented annually so as not to create another set of calf-paths that will need remediation down the road.

5. **Create a registry of collections for your organization.** Include all of your digital content, including collections that you know that you will eventually inherit from your parent organization, in this documentation wherever possible. Tie this registry to your preservation documentation, and include information regarding remediation plans for legacy collections. When you inherit or acquire new collections, make sure that you document and put a price tag on the conversion work.

Recommended Practices for Lifecycle Management of Digital Assets

So, as institutions are implementing a digital preservation readiness program, what details should they plan on documenting to effectively manage the ongoing process of digitization? While we have specific advice that will be offered momentarily, we would first recommend that institutions consider the DCC Curation Lifecycle Model, which provides an overview of the iterative stages involved in curating and preserving digital collections. [2] The model and the series of workshops taught using it as a framework, encourage institutions to think holistically about the entire lifecycle of managing digital assets in terms of related layers of actions and policies. Without recapping this comprehensive model for understanding lifecycle management of data, we will here highlight some additional points of our own. Informed by our experience to date, the MetaArchive Cooperative has documented these practical points that should be carefully defined in an institution's plan.

Live versus Static Media

First, if a collection is deemed of importance to preserve, either now or at some future time, we recommend that the institution go to the effort of storing it on live, spinning discs, not on CDs or other static storage devices. Several of the Cooperative's members had converted the master files of important archival collections to CD-ROM. They did this by archival standards, using "gold" discs to secure their materials. When they were ready to participate in the Cooperative's distributed digital network, they first had to find those discs, load them onto spinning discs, rectify errors and failed media (even gold CDs regularly fail!), and add metadata for these collections. The cost of disk storage is constantly declining at a dramatic rate, and the advantages of having the information available online far outweigh any cost of acquiring and maintaining such storage, which can be accomplished with very inexpensive commodity equipment. Replication through distributed collaborative networks like MetaArchive makes such commodity equipment as reliable as much more expensive SAN infrastructures. As a best practice, we therefore recommend relying on live storage mechanisms whenever possible.

Standardize File and Directory Structures

As described above in the example of Virginia Tech's ETD collection, the file and directory structures used for a collection is of great relevance to its preservation readiness. Most repository systems (whether homegrown, open source, or turnkey) operate on digital assets that are stored on a server file system. Access to the content may be provided by various kinds of indexed databases, but the digital assets themselves are first reposited in the file system through some kind of ingestion workflow. This workflow is often focused from the beginning of digitization projects on the exigencies of throughput rather than organization, because the focus at the beginning of such projects is (understandably) on quickly ramping up production. Unfortunately, because of the calf-path syndrome, the focus all too often does not change, with all the results we have highlighted previously.

We recommend standardizing naming conventions for files and directory structures from the beginning of any project. This will require analysis of the ways that the collection may grow over time, scoping numbering systems that can be parsed automatically, and development of directory structures that can be easily traversed by subsequent harvesting systems. Data structures should also ideally be aligned with item-level metadaa. The point here is to think carefully about the issues involved in automatically processing, wrangling, and migrating the data assets *before* creation, rather than long afterwards. Otherwise, a digitization program will inevitably find itself on the calf-path.

Metadata Discipline

Emphasizing the importance of metadata in digitization processes has become something of a cliché, but a basic understanding of metadata and its purposes is indeed essential for any digitization effort. There are many good overviews available, such as the NISO introduction to metadata. [3] Our experience is that there are still many digitization projects routinely undertaken today with insufficient or nonexistent attention devoted to creation of metadata. These digitization efforts are often undertaken by staffs that know they have inadequate resources for the task at hand and are forced to decide between digitizing materials without adequate metadata and not digitizing anything at all. The most frequent compromise we have seen is in file naming conventions for digital masters that echo labels of archival series, in an attempt to provide scanning archivists with a mnemonic technique for identifying the subject of scanned images for subsequent use in exhibits. This method ultimately fails, as such mnemonic reminders typically do not last beyond the departure of the archivists that used these reminders.

It is always in the best interests of any digitizing staff to create metadata for images, even if this metadata is minimal. There are now so many guides to understanding Dublin Core elements and other metadata standards that there is no reasonable excuse for not imposing basic metadata discipline on a digitization effort. The aim of such a practice should be to associate sufficient metadata with digital assets that they can usefully be accessed and managed by subsequent generations of staff and users. The ideal is implementation of a robust process for assigning metadata by qualified technical experts, but we would be the first here to say that the ideal can be the enemy of the good. Minimal metadata assigned regularly is far preferable to ideal metadata assigned irregularly or not at all. Also, metadata should be mapped in a straightforward, unambiguous, and consistent way to the relevant identifiers of the digital assets (see above comments on consistency in file naming conventions).

Implement a Digital Preservation Viability and Recovery Program

Finally, we strongly recommend that institutions implement a program to assess the viability and recoverability of items

committed to a digital preservation system. Without a program to actively test whether digital assets can actually be recovered from preservation systems, any amount of preparation may be undertaken for nothing. A viability and recovery program should include the following elements:

1. **Assign staff to be responsible for viability and recovery tests.** Unless the activity is officially part of someone's job, it is unlikely to actually take place.
2. **Document the entire process of asset recovery.** Without documentation it is unlikely that the process will really be thought through completely by current staff, and subsequent staff will likely have nothing to guide them in understanding the recovery process.
3. **Recovery tests should be realistic.** Unless the test of asset recovery is a realistic and thorough assessment, you will not really know what to expect in the case of an actual recovery need. Testing the viability of recovered assets includes not just checking to see if the files can be reloaded, but also if they actually display properly.
4. **Conduct periodic tests.** One test is not adequate; the ability to recover specific data assets should be assessed at least annually, and more frequently if possible.

Conclusion

The legacy of problematic digitization practices that we have here termed the calf-path syndrome is a common phenomenon in cultural memory organizations today, at least those that are engaged in digitization activities. The question is probably not whether the syndrome exists in one's organization, but to what degree it exists and to what degree the staff is aware of it and acting to address it. Despite the widespread existence of this syndrome, we think that it can be remediated with steady effort. Many of our recommendations in this paper may seem like obviously needed measures to those not engaged in digitization programs…and too ambitious to those actually involved in digitization. We acknowledge that the steps we recommend do require resources. But the point of digital preservation programs is to avoid the loss of digital assets that may be still more expensive (or simply impossible) to recover. Without taking the measures we recommend, any digital preservation program may be compromised in its ability to actually preserve anything.

The impulse to implement a digital preservation program is not the only trigger event that may alert an institution to the existence of the calf-path syndrome, but in our experience it often provides organizations with the first major opportunity to develop a case for a systematic evaluation of their own digitization practices and collections. This opportunity should be taken; the impulse to leave the calf-path in place for resolution by unspecified future generations is how it persists for so long. We conclude with some final summary recommendations:

1. Admit the calf-path problem exists and needs attention.
2. Isolate calf-paths wherever possible and don't keep following them forward.
3. Implement a digital preservation readiness program and regular "roadway" lifecycle management processes for your new materials.
4. Engage in iterative remediation when you can. Continuing to constrain or totally bulldozing calf-paths may become possible with steady planning.

References

[1] Commission on Preservation and Access and The Research Libraries Group, Report of the Task Force on Archiving of Digital Information, (1996) pg. 7. URL (last retrieved March 27, 2009): http://www.ifla.org/documents/libraries/net/tfadi-fr.pdf

[2] DCC Curation Lifecycle Model, URL (last retrieved March 27, 2009): http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf

[3] NISO, Understanding Metadata (2004) URL (last retrieved March 27, 2009): http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Author Biography

Dr. Martin Halbert is Director for Digital Innovations at Emory University, where he is responsible for researching and leading new library information technology initiative. He has served as principal investigator for grants and contracts totaling more than $5 M during the past six years. He established the MetaArchive Digital Preservation Network (http://www.MetaArchive.org), a consortium of universities acting in concert with the Library of Congress to preserve our cultural heritage as part of the National Digital Preservation Program.

Dr. Katherine Skinner is the Digital Projects Librarian at Emory University and provides leadership for the university's digital projects that are supported through grants or other sponsored funding sources. In this role, she has coordinated efforts involving interdisciplinary interest groups from more than three dozen universities worldwide, including faculty members (in the sciences, social sciences, and humanities), information technologists, librarians, curators, and campus administrators. She is the founding program manager for the MetaArchive Cooperative (http://MetaArchive.org).

Gail McMillan is Director of the Digital Library and Archives and Professor at Virginia Tech's University Libraries. VT set the national and international standard for ETDs and she has played a significant role in this initiative since 1995, including developing software that is used at universities throughout the world to manage ETDs. Under her direction, DLA is responsible for all aspects of ETDs from students submitting them online to user access and long-term preservation.