Getting ETDs off the Calf-Path: Digital Preservation Readiness for Growing ETD Collections and Distributed Preservation Networks

Martin Halbert, Katherine Skinner[1]; Gail McMillan[2]

[1]Emory University; Atlanta, GA. [2]Virginia Polytechnic Institute and State University; Blacksburg, VA.

mhalber@emory.edu, kskinne@emory.edu, gailmac@vt.edu

ETD repositories often start with very idiosyncratic and ad-hoc beginning data storage structures, 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. ETD repositories also tend to grow in an effectively unbounded manner over time. These early idiosyncrasies and unbounded growth can subsequently cause enormous problems in systematic efforts to digitally preserve content of growing collections. The most effective preservation strategies incorporate pre-coordinated replication of content in distributed and secure locations; such replication strategies become increasingly difficult when the content is stored using irregular practices in directory structures, metadata, and file naming conventions. This paper will address "Calf-Path" problems by providing practical guidelines, suggestions, and recommendations for ETD repositories. These recommendations are informed by five years of experience in operating the MetaArchive Cooperative, a distributed digital preservation cooperative of cultural memory organizations which has grappled with standardizing transfer mechanisms and developed cost/effective strategies for distributed preservation of ETDs based on the LOCKSS open source software. In the course of the past six years the members of the MetaArchive Cooperative have identified a series of best practices for digital preservation readiness. 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. This paper will document relatively simple principles and guidelines for such programs that can greatly improve the subsequent likelihood of implementing successful distributed digital preservation programs.

Introduction

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.

-Sam Walter Foss, "The Calf-Path"

University electronic thesis and dissertation (ETD) programs have rapidly grown during the past two decades, and have now become an essential service provided to academic communities by libraries. Yet, 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.

When managers decide to establish a digital preservation program for their institutional ETD collections the ad-hoc procedures that initiated the ETD service may cause enormous problems. This paper will speak to 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 that has preserved ETDs from its inception. 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 ETD repository programs per se from true digital preservation programs, as we consider repository systems a means of managing workflow and access to digital collections rather than a full digital preservation strategy.

How ETDs get on the Calf-Path

The inception of ETD preservation programs in the late Twentieth Century was a major evolution of the traditional thesis and dissertation deposit programs operated by university libraries. It enabled libraries to bring to bear all the strengths and affordances of network infrastructures for the process of managing theses and dissertations archiving. Yet, most cultural memory organizations (like many other institutions in society) lack experience in what Don Waters and others have called "Deep Infrastructure" [1]. Many ETD repository efforts have very humble and informal beginnings with limited resources and staffing, but nevertheless may gradually grow over time to produce and process very significant bodies of content and volumes of material.

In this paper we will discuss a variety of findings concerning ETD repository programs in more detail through a case study from MetaArchive. Such programs tend to grow in an effectively unbounded manner over time, and 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 online 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 an ETD repository through special one-time campus funding for a pilot project. As part of the project, the organization creates a directory structure and naming conventions that make sense within the narrow confines of the pilot. A few years later, different individuals within the same organization (but perhaps from the archives division as opposed to the systems group) might gain funding to retrospectively digitize older theses in a particularly disciplinary area. 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 ETD 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. Eventually a comprehensive program is blessed and (inevitably?) institutionalizes the entire assembly of ad hoc practices that grew up over time without ever "straightening the calf-paths" that are now formalized as "roads" of workflow.

In this way, multiple "calf-paths" may be created in universities by groups with different professional practices, one by archivists, another by digital librarians, and yet others by records managers. They all belong to the same organization and share a need to preserve the various collections of institutional intellectual output 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. A case study of electronic theses and dissertations (ETDs) is a practical example of a type of organizational content that often 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 open access for the long term 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, as well as participants in the ETD-L. The results of this survey were reported at the 11th International Symposium on Electronic Theses and Dissertations held in Aberdeen, Scotland, in 2008. [2]

Among the responses to the 18 questions, we learned that the file formats received for ETDs are 85% PDF; 20-30% are JPG, WAV, GIF, HTML, MOV, AVI, and MP# among others. However the survey revealed that only 27% of the institutions have formal preservation plans for their ETDs. Nearly 75% indicated that they would share preservation responsibilities by running a secure server for PLN. The MetaArchive Cooperative and NDLTD, therefore, began to plan a pilot project to examine the practical issues involved in a collaborative replication strategy for digital preservation of ETDs. These findings (and others) are 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 the goals even when a specific plan is lacking. Regular 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, access, and storage are established.

Directory structures and file naming conventions for ETD collections are frequently created without considering their impact on the collections' preservation readiness. As a result, ETD collections often grow almost arbitrarily, seemingly structured but lacking the logic that favors long-term 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 that are logically named. 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 the next SIP be prepared, 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 with established ETD initiatives. In these cases, if there is not yet time for remediating older files, at least the files created in the future can be geared toward long-term access and preservation readiness.

ETDs: Recommendations and Best Practices

Effectively organizing an ETD collection for preservation readiness requires, among other things, 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 beneficial, for example year/month would be 200801, but not 2008-January or 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. Not only is this fundamental to LOCKSS but it keeps future calf-paths from developing if there is a loss of both institutional memory and lack of written procedures because the directory structure is obvious, logical, and easy to continue.

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 ingrained calf-path for years? Such older collections 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: ETDs at Virginia Tech

Because Virginia Tech has a 12-year history of ETDs, having begun requiring them on Jan. 1, 1997, and as a founding member of the MetaArchive Cooperative, we discovered early on that a variety of URN (Uniform Resource Names) conventions had been used from 1996-1999. Ergo, a calf-path developed to a hodgepodge of URNs. Students who submitted their works through the ETD_db prior to the major software upgrade in 2000, however, were at least assigned unique identifiers. The URNs were not consistently structured from today's point of view and they went through several iterations. For example, a dissertation labeled etd-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 complexity of this static collection is best served by atypical plugin rules that *exclude* anything that matches the “proper” (i.e., post-1999) URN structure and places it into an "early VT ETD" collection.

Since 1999 the established and recommended naming convention for ETDs follows the consistent format *etd-mmddyyyy-tttttt* and is based on the timestamps received when students begins to submit their ETDs to the Graduate School for approval at the university level (having already met with committee/department approval). All ETDs submitted after 1999 become part of the ETDs@VT collection, which is subdivided into annual archival units.

Some ETDs from the calf-path era also include unpadded months and days as well as two- and four-digit years but these are correctly harvested into AUs by year. Therefore, anything that does not match the recommended file naming structure becomes part of the collection of early ETDs. At Virginia Tech this collection is named “ETDs@VT-pre2000unsorted” 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 anything that does not follows the recommended format.

Scanned (versus born-digital) theses and dissertations follow the recommended file naming convention based upon the timestamp for the digitization date, not the original date on which they were approved. This allows the calf-path collection to remain static and unchanged when older theses and dissertations are scanned. Instead, they are ingested through the existing plugins and preserved in the AUs with the born-digital ETDs. Using a time stamp as a naming convention works well for preservation purposes because it continues to generate consistent, logical, and unique names for every electronic thesis and dissertation, whether born digital or scanned.

Alternatively, it would not be very complicated to programmatically generate URNs for scanned theses and dissertations based on their completion date because this information exists in the host institution's MARC bibliographic records. ETDs are usually assigned Library of Congress call numbers that include dates. For example, the dissertation with the call number LD5655.V856 1994.L556 is based on the formula: Institution number--LD5655, dissertation designation--V856, year--1994, Cutter number--L556 (based on the author’s last name, Limoges).

However, batch processing involves not only file naming but also pulling volumes from possibly multiple locations (e.g., main library and remote storage), arranging them in order, maintaining their order, accurately deriving the file names from the MARC records, and linking them to the appropriate files. This may be very cumbersome and inefficient in terms of preservation goals. The MetaArchive model separates the preservation function from access, which remains with the host institutions. If it becomes necessary to rebuild the database of digital theses and dissertations (both born-digital and digitized), access, arrangement, and display is determined by each institution and is external to the purpose of the PLN.

This case study 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, a reasonable strategy is to isolate it with special signage and create a path directly to the intended outcome.

From Calf-Path to Direct Route

Assuming an organization has recognized a calf-path in its midst, we believe that an initial helpful activity is to develop a digital preservation readiness program. The most effective preservation strategies incorporate pre-coordinated replication of content in distributed and secure locations. Such replication strategies become increasingly difficult to implement when the content is stored using inconsistent practices in directory structures, metadata, and file naming conventions; in short, when digitization efforts are trapped on a calf-path. As universities develop in digital preservation strategies, what becomes apparent is the need for clear guidelines to help them structure collections for preservation readiness. During the last six years the MetaArchive Cooperative has worked with institutions to articulate principles and guidelines for such programs that can greatly improve preservation readiness. We feel that these best practices can benefit new ETD initiatives as well as 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. This applies not only to ETDs but all digital collections. But how do you establish a preservation readiness 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 experience. It applies to any digital program including ETDs.

1. **Start with a shared programmatic vision.** The key word here is "shared." From assessment to publication, representatives from across the organization should design the program. It should also have the buy-in of the organization's head. For ETDs, this would include be the dean (or equivalent) of the library and the graduate school(s).
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 may become involved in the future. Document whether the collections need to be remediated to fit this new set of best practices, even if there are no resources initially 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 staff through discussions, memoranda, and documented guidelines.
4. **Review your vision and best practices annually.** Keep the documentation alive through dating its production and scheduling annual reviews by the spectrum of those involved in the digital initiative. All of us know that the digital landscape is ever changing in the 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 later.

5. **Create a registry of digital collections for your organization.** Include all of your digital content in this documentation wherever possible, including collections that you know you may eventually inherit such as bound theses and dissertations that may be scanned at some future time. 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

As institutions are implementing a digital preservation readiness program, we 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. [3] 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 highlight some additional points of our own and relate them specifically to ETDs. Informed by our experience to date, the MetaArchive Cooperative has documented these practical points that should be carefully defined in an institution's preservation plan.

Live versus Static Media

ETDs record the research and scholarship of the institution's graduate students and, consequently, its faculty, and are deemed import to preserve. Therefore, we first recommend that institutions go to the effort of storing them on live, spinning discs, not on CDs or other static storage devices. Several of the Cooperative's members previously 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 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 online 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 case study of Virginia Tech's ETD collection, the file and directory structures used for a collection are 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 a digital initiative 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 the adverse consequences we have highlighted.

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 metadata. The point here is to think carefully about the issues involved in automatically processing, wrangling, and migrating the data assets *before* submission of the born-digital ETDs or scanning bound theses and dissertations, rather than long afterwards. Otherwise, a digitization program will inevitably find itself on the calf-path.

Metadata Discipline

Emphasizing the importance of metadata has become something of a cliché, but a basic understanding of metadata and its purposes is indeed essential for any digital initiative, not just ETDs. There are now so many good overviews available, such as the NISO introduction to metadata [4] as well as guides to understanding Dublin Core [5] elements and other metadata standards, that there is no reasonable excuse for not imposing basic metadata discipline. 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 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 (e.g., see above comments on consistency in file naming conventions).

It is, however, our experience that there are still many digital initiatives routinely undertaken with insufficient or nonexistent attention devoted to the creation of metadata even though we know that “Good quality metadata is easy to provide at the point of creation but usually difficult, expensive or impossible to discover retrospectively.” [6] Fortunately, it is standard practice for university libraries to catalog theses and dissertations, thus ETDs can have ready-made metadata derived from MARC [7] or metadata can be derived from the descriptive elements authors enter when they submit their ETDs.

In 1998 the NDLTD established and continues to maintain a metadata standard for theses and dissertations called ETDMS [8], and it created a crosswalk between these metadata elements and the MARC fields. Based on the Dublin Core, ETDMS has additional fields specific to ETDs, and it can handle metadata in many languages, including a single work that has multiple languages. While ETDMS may not have the status of an official international standard metadata schema, it has a well-established user base in the US and outside including Canada where it is the standard of the (national) Library and Archives Canada. There are also versions of ETDMS for German, French, and Brazilian ETDs. OCLC, among others (e.g., Scirus, VTLS), harvests ETD metadata via OAI-PMH from around the world to create a union catalog of harvestable ETD metadata.

Preserving ETDs means not only preserving the files that comprise these works, but also keeping the metadata that describes each work. The key is to preserve both so that if it becomes necessary to restore an ETD, you will have both the work itself and its identifying information, i.e., its metadata.

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 early digital initiatives that lead to problematic digitization practices that we have here termed the calf-path syndrome is a common phenomenon in institutions today that are engaged in ETD initiatives and 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 may seem like obviously needed measures to those not engaged in digital initiatives, and too ambitious to those long-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 digital collections and their digitization practices. 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.

Therefore, some final summary recommendations are:

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 lifecycle management processes for new materials.
4. Engage in iterative remediation when possible. Continuing to constrain or totally bulldozing calf-paths is possible with good planning and steady remediation.

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] Maintaining the Light: Distributed Digital Preservation of ETDs. Proceedings of the 11th International Symposium on Electronic Theses and Dissertations, Robert Gordon University, Aberdeen, Scotland, June 5, 2008. URL (last retrieved May 12, 2009): http://www.rgu.ac.uk/files/ETDPreservSurveyPaper.ppt

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

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

[5] The Dublin Core Metadata Initiative developed and maintains the interoperable online metadata standards that support a broad range of purposes and business models, including educational efforts to promote widespread acceptance of metadata standards and practices. URL (last retrieved May 20, 2009): http://dublincore.org/

[6] ETD Guide. NDLTD. URL (last retrieved May 14, 2009): http://en.wikibooks.org/wiki/ETD_Guide

[7] MARC is the acronym for MAchine-Readable Cataloging. It defines a data format that emerged from the United States’ Library of Congress-led initiative. It provides the mechanism by which computers exchange, use, and interpret bibliographic information, and its data elements make up the foundation of most library catalogs used today. URL (last retrieved May 20, 2009): http://www.loc.gov/marc/faq.html#definition

[8] ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations. NDLTD. URL (last retrieved May 14, 2009): http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html/