

Chapter 5: Content Selection, Preparation, and Management

Gail McMillan (Virginia Tech)

Rachel Howard (University of Louisville)

OVERVIEW

This chapter covers recommendations for policies regarding the selection, preparation, and management of digital content preserved in a Private LOCKSS Network (PLN). It provides best practices for organizing and data wrangling collections of both scanned and born-digital materials. These best practices may also be applied more broadly to many distributed digital preservation (DDP) initiatives, as the fundamental principles of collection organization will be similar across such approaches. These policies and best practices may affect existing digital collections as well as the planning of future digital projects. This chapter also describes some of the specific techniques and technical steps that participants in a PLN will need to follow as they ready their collections for ingest by the LOCKSS software. For further best practices regarding ingest, monitoring, and recovery of digital content preserved in a PLN, please see Chapter 6.

CONTENT SELECTION

We cannot preserve everything digital, nor would it be particularly useful to do so. Digital content, just like print and object-based content needs to be identified, collected, organized, prioritized, and preserved.

A key difference between traditional and digital preservation is that digital preservation needs to start early enough in the digital object lifecycle for it to be viable. While a brittle book or sticky acetate tape may be salvaged for reformatting after the damage has begun, even slightly corrupted digital materials are not as easily rescued. They cannot be subject to benign neglect, or be created and then ignored for decades, since their formats, software, and/or hardware may degrade, or become obsolete and their storage locations obscured thus rendering them inaccessible over a relatively short span of time.

An early step toward preserving digital files is to identify one or more experts within an institution to determine exactly what content should be preserved. These experts are generally librarians, archivists, curators, and the like who are knowledgeable about digital formats, issues of scope, copyright status, as well as the risk factors associated with digital archive content. These elements are discussed in greater detail below.

Formats

Decisions regarding what formats to preserve may be made at the network level or the local level for any distributed digital preservation solution. Some solutions will dictate format as part of the software itself and others will preserve any file regardless of its format. For example, the LOCKSS software is format agnostic, meaning it will accept any computer file in any format. Therefore, contributing institutions (or content contributors) that are planning to host a PLN in particular have two options regarding the file formats they accept and ingest for preservation. They may establish criteria for participation that are format-based, or they may leave the decision about what formats are worth preserving to individual content contributors in their network. In other words, PLNs can be as broad or as narrow as they choose when establishing format-based criteria for participation, and need not be constrained by format decisions made locally at their contributor sites.

Whether the DDP network (PLN or otherwise) or the content contributor makes the decision regarding what formats it will preserve, there are emerging community-wide best practices that can guide the decision-making process. The main consideration, of course, is that some formats will be more accessible than others in the long term. Rare or esoteric formats may require more ongoing investments by content contributors in order to maintain their viability over long periods of time. Bit-level preservation should ensure that content is being responsibly preserved and managed in the interim of any major advances in format validation and migration. One such resource that is advancing, and may prove useful to DDP networks in the future, is the Unified Digital Formats Registry (UDFR), which is an international alliance that is creating a format registry to identify formats approaching obsolescence, and new successor formats that may be viable for migration.¹

There are also several publications currently recognized in the library community as excellent guides for evaluating file formats for long-term preservation. Two in particular provide specific criteria for determining the durability of any file format, recognizing that the future will call for migration to new formats, emulation of current software on future computers, or both.

The first of these guides is *Sustainability of Digital Formats: Planning for Library of Congress Collections*.² This publication describes seven factors that influence the feasibility and cost of preserving any particular file format. In addition, the article discusses quality and functionality factors to consider, as well as the need to find a balance between best practices and the realities of donated digital objects that can be quite varied.

The second publication is an article from the National Library of the Netherlands (NLN) that identifies seven sustainability criteria. Titled "Evaluating File Formats for Long-term Preservation," the article provides quantifiable measures for each criterion, acknowledging that pragmatically not all criteria are equally important.³ The weights that NLN applies can be adjusted for other organizations, as they are based on a combination of local policy, digital preservation literature, and common sense.

These two examples of file format selection guidelines for preservation recognize that there are a number of potentially competing sustainability factors that must be weighed on an individualized basis. They help to provide readers with an awareness of the range of file formats, some of which can be virtually guaranteed to be sustainable, some of which are likely to be sustainable, and some of which the level of sustainability is as yet unknown.

Another format consideration for DDP networks and their contributors is the determination of what constitutes a master file, and whether this master version is the only one to preserve, or if it is also desirable to preserve derivative (i.e., access) versions. Some consider the master file to be the original scan, original video-capture, or the first digital form of an object (born-digital or digitized). Others consider it to be the richest version of the file that is in use; for example, the master file of a scanned book page would not necessarily be the raw scan or capture, but rather the uncompressed file that has been cropped, rotated, and color corrected for production purposes. Some consider the version of the digital object that best represents the original content to be the

best version to preserve. Organizations such as the Digital Library Federation (DLF) have developed standards for elements of a digital master registry, such as the recommendations found in their "Registry of Digital Masters Record Creation Guidelines."⁴ Each content contributor at the network level should carefully consider future use scenarios and make preservation decisions accordingly. It is highly recommended to preserve both master and derivative files so long as the network has (and the contributing institution can afford) adequate storage space.

Scope

The decision to form a DDP network is often made based on a shared collection focus that is common among the prospective contributing institutions. The common areas may fall under any number of criteria, such as topical content (e.g., the MetaArchive of Southern Digital Culture archive), genre (e.g., electronic theses and dissertations, or U.S. government documents), format (e.g., data sets), and even location (e.g., statewide initiatives such as the Alabama Digital Preservation Network). This common ground among disparate content contributors may be determined through informal networking or through the analysis of collection data, such as that collected by survey (e.g., the "Electronic Theses and Dissertations Preservation Survey" conducted by the MetaArchive Cooperative in collaboration with the Networked Digital Library of Theses and Dissertations).⁵

It may seem obvious to preserve subject-specific digital materials in a DDP when all the content contributors have holdings in similar fields. However, if inclusiveness is a priority, contributing institutions can select shared subject matter, but define the scope very broadly to include materials that may not be immediately obvious. For example, the MetaArchive Cooperative decided to define its focus of Southern heritage very broadly for its Southern Digital Culture archive, allowing the inclusion of less subject-driven materials such as university archives in the geographic South, in addition to the more traditional subjects of this region, including the Civil Rights Movement, the railroad industry, slavery, and the Civil War.

Decisions around the comprehensiveness of scope may be achieved through other means as well. The content contributors may choose to collect materials relating to a specific topic, but restrict it by time period, geographic region, or genre. It may be important to reserve leeway for contributing institutions to include

materials that might not strictly fit the criteria of the network but, nevertheless, warrant preservation. The founding principles of a specific DDP network can determine how narrowly or broadly to define the scope of the archive.

The collections of the contributing institutions need not be homogeneous to create a successful network; however, by establishing a common scope, a DDP network may develop a common sense of purpose among the content contributors that are jointly investing in preserving each other's collections.

After determining the scope of the preservation archive, the next step is to document what content is on-target, and how fluid the definition is, and if necessary, to consider circumstances for expanding the scope. These early decisions made and documented by the members of a DDP network may not only affect the rate at which the network grows, but also its ability to attract new members.

Copyright Status

Copyright, which will be discussed further in Chapter 8: Copyright Practice in the DDP, must be considered when establishing a DDP network and selecting digital materials to preserve.

Many preservation efforts conflate maximizing short-term access (i.e., high availability) with long-term access (i.e., preservation). High availability entails adopting strategies for ensuring that content is constantly available to the public. It also mandates that content is free of copyright and intellectual property constraints through the use of appropriate licenses or permissions owned by the contributing institution.

A DDP network may be an open archive, or it may reside somewhere on the spectrum from dim to dark archive. That is, it may be open to only the contributors' servers for ingesting (dark archive); it may be open to specified users, such as the contributing institutions' communities (dim archive); or it may provide unrestricted access (open archive). This status will determine whether contributors will focus solely on long-term preservation issues, or some combination of preservation and public access issues.

- **Open PLN Archives:** CLOCKSS (Controlled LOCKSS) is a not-for-profit, community-governed, alliance of research libraries and publishers.⁶ Though somewhat different from many of the PLNs explored

throughout this book due to its journal content targeted for preservation, CLOCKSS is an excellent example of an open archive. For example, when a trigger event has occurred and the digital content is no longer available from a publisher, one of the participating institutions will move the content to a hosting platform and the impacted preserved content will be made available without charge to the world.

- **Dark PLN Archives:** Neither the MetaArchive Cooperative⁷ nor the Alabama Digital Preservation Network (ADPNet)⁸, as dark archives, has a public access component at the present time. Preservation and access, though united in their goals, are considered two separate functions. Only content contributors in the PLN have access to the collections, and this access is restricted to ingesting collections into preservation caches and to restoring digital content to the contributor (i.e., not to view or use the content). The contributing institution determines whether access is provided or not.
- **Dim PLN Archives:** The original LOCKSS public network provides preservation and access to content governed by the legal or license agreement associated with that content. For example if a subscription publisher limits access to a range of IP addresses, access to that publisher's LOCKSS preserved content is limited to the same range of IP addresses. Government documents are also preserved in the LOCKSS system. This content is not subject to any further access restrictions either from the publisher or from the LOCKSS system.

Whether or not the preservation network accommodates public access to the preserved content, each member institution must be responsible for implementing appropriate standards for addressing copyright, intellectual property, and issues related to content that has been contributed. Content contributors bear the responsibility for determining ownership and their rights to preserve the content prior to submitting it to a DDP network. Compliance with laws – including the use of exemptions set forth within U.S. Code Title 17 (copyright law) in sections 107, 108¹⁰, and elsewhere, and permissions through deeds of gift or other clearances is an obligation of each institution in the PLN. International institutions

must similarly address intellectual property and copyright laws. Rights should be documented in the collection-level metadata.

Preservation networks rely on a great deal of trust, including the trust that contributing institutions are not violating copyright law when sharing their digital files, even for preservation purposes. Trust needs to be formalized in a legal agreement indicating that contributors represent and warrant that, to the best of their knowledge, they are not contributing content to the preservation network that would infringe the rights of others. Each contributor should also certify that it holds sufficient rights to authorize the DDP network to use the content in a manner consistent with the requirements of a multi-cache preservation strategy, whether it is a dark archive or one that provides some level of public access.

The Membership Agreement for the MetaArchive Cooperative is an example that covers these formal issues with appropriate terminology and legal language.¹¹

Risk Factors

One of the major concerns when pursuing digital initiatives is the fear of loss due to many potential factors, including natural disasters, human errors, fires, floods, power surges, and more. However, previous worries about unstable media and hardware obsolescence have been greatly reduced after more than two decades of providing digital media to library constituencies. An excellent grounding in these issues is available in the Council on Library and Information Resources' 2000 publication, *Risk Management of Digital Information: A File Format Investigation*.¹² It outlines a variety of factors that might put a digital collection at risk, and supplies a pragmatic approach to assessing risks of digital collections with its "Risk-Assessment Workbook." The DRAMBORA assessment tools likewise supply institutions with a workbook approach to risk assessment and management.¹³

In addition to the safe harbors that are created in a DDP network, it is also important that the contributing institutions in the network make decisions based on long-term access goals [or strategies], not just current technology. Once the content contributors have agreed on the risk factors for their collections, they can assign priorities for ingestion into the network using risk rankings. Because not all content can be ingested simultaneously, and not all content may be worth preserving, each DDP network may wish to set risk guidelines to prioritize content for ingest. They might also review

where files are stored, and on what type of media. For example, large digital master files may exist solely on external media, such as compact discs (CDs). In order to be ingested by the other members of a PLN, the files would need to be transferred onto a web server and arranged into archival units (AUs). Files stored on servers are likely to be safer in most cases than those on offline storage media such as CDs and DVDs. Finally, it might consider whether a file has been backed up, and if so, whether those backups are tested regularly.

With these issues in mind, consider the risk levels adopted in 2004 by the MetaArchive Cooperative as it launched its Southern Digital Culture archive. In 2009, the Cooperative still uses these guidelines for this archive, and has extended them so that they may also be applied to new archives established by the Cooperative:

1. **Extreme Risk:** No one is responsible for preservation. No other copies of the digital content are preserved. No regular backups or data migration.
2. **Significant Risk:** Responsibility under discussion, but no copies of the digital content are currently being preserved.
3. **High Risk:** Only one backup of digital masters on CD-ROM. No regular backups or data migration.
4. **Moderate Risk:** Some danger that collection backups might be lost in the future.
5. **Low Risk:** Copies are backed up regularly with a long-term maintenance plan in some other trusted digital archive.

CONTENT INGEST PREPARATION

Organizations create digital collections as part of their ongoing work, but often ignore or set aside long-term planning, which results in idiosyncratic and ad-hoc data storage structures. Such early idiosyncrasies can become embedded in these collections' data structures, upon which digital infrastructure and management workflows continue to be built. Such infrastructures may cause prodigious problems during systematic efforts to preserve the content of these (static or growing) collections.

This section outlines two important components, one that is broadly applicable to DDP networks and one that is specific to PLNs. First, it will outline how to prepare a collection to be programmatically ingested into a DDP network, and then it will specify how to initiate content ingest within a PLN. It provides examples of up-front planning with long-term preservation in mind, including clearly defined and documented collection data structures. It also suggests remedies for collections that evolved with little or no direction.

Content ingest requires the following elements:

- Accessibility (for PLNs, this must occur using the Web)
- Organizing Collections
 - Data Wrangling
 - Metadata Creation
- Defining Archival Units for a PLN Solution
- Manifest Page Creation for a PLN Solution
- Plugin Creation for a PLN Solution

Each of these elements is described below in greater detail. For more information about basic DDP and PLN architecture, please see Chapter 2: DDP Architecture. For additional details on preparing content for ingest into a PLN network, please refer to Chapters 6: Content Ingest, Monitoring, and Recovery.

Accessibility

For any DDP network to ingest/harvest content, it must first be made accessible to that network. This may occur through a submission process in which the contributing institution sends files to a central location or it may happen through web-based harvesting or other mechanisms.

For example, in order for content to be ingested into a PLN, it needs to be web-accessible via HTTP (Hypertext Transfer Protocol) or HTTPS (secure HTTP). When access restrictions are in place (e.g., only constituents from the contributing institution have access), a list of specific preservation members' IP addresses must be added to the web server's firewall configuration to enable ingest by the authorized PLN institutions. For more details on this

configuration please see Chapter 7: Cache and Network Administration for PLNs

Organizing Collections

For preservation and life-cycle management purposes, a digital collection and its content should be clearly arranged, defined, and described. When beginning a digital initiative, it is wise to consider what might be necessary for both programmatic capture and online user access, such as hierarchical arrangement and logical file naming (see the section on “Content Management” below).

When creating a new digital collection, it is highly recommended that an institution organize it into a methodical or hierarchical file structure. For example, an Electronic Theses and Dissertations (ETD) collection may require a new directory for each submission year. Digitized special collections could follow the same organizational structure as the physical collection, which often has a hierarchy of folders within a series. Naming conventions should include logical labels for each folder in each series. The series can be organized by subject, as well as chronologically or alphabetically. Even when only a portion of a collection is digitized, a complementary file directory structure should be established to better manage the long-term preservation of the digital items. This practice will avoid the creation of a directory that is a hodgepodge of files. Other logical arrangements could resemble a business organizational structure, a genealogical family tree, or a calendar of events. Documenting any policy that is developed helps to ensure its understanding and usage by future digital collection managers.

For the purposes of the MetaArchive Cooperative’s PLN, a collection is defined as the aggregated content to be preserved under the banner of one collection-level metadata record, which is entered in the Cooperative’s conspectus database. (see the section on “Metadata,” below.) It may differ from the original analog or digitized collection because the entire collection may not be digitized or digitally preserved due to copyright, risk, or other reasons.

Data Wrangling

It is not atypical to encounter existing digital collections that were created without forethought, resulting in rather haphazard collections that are not preservation-friendly. Data wranglers

alleviate these problems by wrestling the digital objects into discernable units.

When associated with a PLN, data wrangling refers to the strategic rearrangement of digital collections so that the path to them can be logically defined for programmatic access. In order for the content to be ingested into the PLN (which uses web-based mechanisms for this ingest process), some data wrangling may be required to assemble the files into a coherent order (or to identify their location) and to describe the collection clearly and thoroughly for effective future access. This effort has been particularly necessary for older collections established in the early days of the Internet, when making them electronically accessible was often rushed and not approached in a strategic, long-term manner.

Data wrangling may entail moving and rearranging master files and metadata into directories and folders corresponding to newly created file directories for the collection and its sub-collections (or, in the PLN context, its Archival Units (or AUs, as described in “Defining Archival Units” below) This inevitably leads to discoveries of missing, mis-numbered, duplicated, substandard, or corrupted files, as well as insufficient metadata. Identifying and correcting these errors will aid not only in preserving the digital assets, but also in providing both short- and long-term access to them.

Qualified staff members who know the custodial history (provenance) of the materials to be preserved should make the decisions about arrangement and description. However, university members of PLNs have found student employees to be effective data wranglers, preparing collections for ingest by moving files or creating virtual collections. As described further in Chapter 4: Organizational Considerations, data wranglers may also write plugins and manifest pages to permit digital content to be ingested by LOCKSS. This preservation work is sometimes analogous to processing physical archival collections, which must be arranged, inventoried, and re-housed before they can be accessed.

Metadata

Digital preservation depends in large part on ascribing effective metadata (structural, technical, and descriptive) to objects and collections. DDP networks, including PLNs, have to make choices about what metadata standards they wish to employ, and at what level: the network or contributing institution. That metadata aids preservation is an uncontested principle; however, metadata

standards can become a barrier to entry for potential network participants. Each DDP solution must weigh the pros and cons of such metadata standards as PREMIS and METS, and must determine what level of standard best suits the preservation needs of its member institutions.

For example, the MetaArchive Cooperative has found that collection-level metadata is an essential tool for its preservation network, as it facilitates tracking and maintenance of the content. Contributors with backgrounds in archives, systems, cataloging, and digital libraries can be helpful in fully describing collections in ways that are meaningful to both the contributing institution and to the network monitoring process. It is important that they not only have knowledge of the collection, but also understand the preservation goals and functionality of the PLN. Detailed information about each ingested collection also facilitates network management and assists with various access-related issues, including disaster recovery, where a contributor needs to use the preserved digital content to rebuild its local collection. The Cooperative does not, however, require its contributing institutions to limit their preservation activities to those collections that have item-level metadata in any particular schema, as the differing practices of its member institutions means that any such requirement would necessarily limit the preservation of their collections.

There are a number of excellent existing schemas that can be used or adapted to meet a DDP network's collection-level metadata needs, including the following:

- BCR CDP's Dublin Core Metadata Best Practices¹⁴
- Dublin Core Collections Application Profile¹⁵
- UKOLN Research Support Libraries Programme (RSLP) Collection Description Schema¹⁶
- IMLS DCC Collection Description Metadata Schema¹⁷
- PREMIS Preservation Metadata: Implementation Strategies¹⁸
- MetaArchive Collection-Level Conspectus Metadata Specification¹⁹

Each schema contains standard elements for library and archival description, such as title or creator. Some metadata elements in

these schemas are tailored specifically to digital objects, such as MIME format. As a DDP network considers which elements to include in its schema, it should think about how it wants to record, and in what order the materials will be ingested, as well as information regarding accrual, or how often a particular collection is updated. These elements are important for ingest and for storage projections. Depending on the needs of the DDP network, the collection description can require controlled vocabulary (e.g., Library of Congress Subject Headings) or code (e.g., ISO 639-2 language code).

Continuing with the MetaArchive Cooperative's example, the Cooperative determined that there are eight principal categories of metadata elements:

1. Descriptive data illustrates or explains the collection.
2. Uniform resource identifiers (URIs), uniform resource names (URNs), and unique identifiers locate the collection.
3. Coverage places the collection in space and time.
4. Accrual information anticipates the growth of the collection.
5. Data description provides formats, sizes, languages, etc.
6. Rights and ownership elements document intellectual property and provenance.
7. Related resources inform about associated collections.
8. Ingesting information provides data necessary for the ingest process.

In order to identify, ingest, and track the collections of a DDP network, each contributor may record collection-level administrative and descriptive metadata in a DDP-specific database (in the case of the MetaArchive Cooperative this is the conspectus database, which is freely available to other PLNs). This database describes the breadth of the DDP network through network-wide and institution-level views. Each collection, which in the PLN arena may be comprised of one or more AUs, has one corresponding metadata record in the database. The database

should provide metadata versioning support to track collection changes.

The conspectus database designed by the Cooperative interoperates with the LOCKSS title database, providing relevant information in an XML dialect of RDF. The title database contains the XML parameters that tell the LOCKSS daemon three central things: 1) where to find plugins as signed jar files, 2) the location of archival units, and 3) the list of IP addresses for caches participating in a network. The consistent use of XML makes it easier for the conspectus database to generate the title database as well. To this end, the conspectus also records metadata that is required for ingest by the LOCKSS software: the plugin name, plugin parameters (where used), and the base URL of the collection.

The Cooperative also recommends preserving local item level descriptive, administrative, and structural metadata for the digital objects in the collections wherever such metadata exists. The metadata should be in a sustainable format such as unformatted (ASCII) text or XML, and should be ingested by the PLN along with the collections they describe.

Defining Archival Units for a PLN Solution

As described in more detail in Chapter 6: Content Ingest, Monitoring, and Recovery, PLN ingests are conducted through guided crawling, which is much more exact than typical web-spidering methods. PLN ingests target specific collection components based on their Archival Units (AUs) — which are the collection boundaries established by the content curator before a given collection is slated for ingest.

AUs are the building blocks of a LOCKSS collection. An AU is a cohesive and logical aggregation of content by topic, format, file size, or file location that is intended to divide a collection into discrete groupings (typically between 1 GB and 20 GB in size) for ingest into the PLN. For example, each AU of a collection of digitized yearbooks might comprise a single volume, or a collection of ETDs could have AUs for each year's theses and dissertations.

AUs for large digitized manuscript collections may correspond to the hierarchical folder arrangement. That is, the files and metadata may be organized into record groups, boxes, folders, sub-folders, and items. For a collection of photographs, an AU might be the entire collection, or it might be a folder containing the digital

masters for the collection. If size and organization permit, an AU can encompass all items in an entire record group.

Examples of AUs include:

- One volume of an e-journal
- One year of ETDs
- One decade of scanned yearbooks
- One folder of archival TIFF images or sound files

Manifest Pages for a PLN Solution

Each AU must have a manifest page, which serves as a starting point for ingest, and a statement granting permission to LOCKSS to ingest the AU. The manifest page is usually a normal HTML page, and must link (usually indirectly) to all the content that should be included in the AU.

The permission statement is usually contained on the manifest page, but it may be located anywhere on the same host as the content to be ingested. Either the following statement: "LOCKSS system has permission to collect, preserve, and serve this Archival Unit", or a Creative Commons license, is acceptable. The statement need not be visible to users, e.g., it can be placed within an HTML comment.

Best practices for manifest pages include:

- Making sure AUs are properly accounted with an individual manifest page, or a collection-level manifest page.
- A manifest page should avoid, when at all possible, attempting to encapsulate a complete list of files that are to be ingested. It should instead point to the location of the AUs.
- Although not required by LOCKSS, it will assist long-term preservation efforts if each manifest page contains the name of the collection, the institution, and a contact name/address, and is updated to reflect changes in this administrative information as well as changes in the AUs.
- Manifest pages should contain a short description of the structure of the collection, such as where to find

metadata, the naming conventions used in filenames, and how the AUs relate to the site structure.

- A collection's manifest page should contain a link to its collection-level metadata description.

Plugins for a PLN Solution

A plugin provides information to LOCKSS about a collection or a group of similarly structured collections to tell it how to collect and audit the content. A plugin is a small block of XML which, when given a set of AU-specific parameter values (such as base URL and year), defines the URL of the AU's manifest page.

LOCKSS has a plugin tool²⁰ and a plugin tool tutorial²¹ freely available online. In addition, the MetaArchive Cooperative has created a Plugin Standards Checklist to guide the plugin creator through the Java coding decisions.²² Virginia Tech has also produced a plugin tutorial that contains a case study on ETDs.²³

As plugins are under development for a collection, they should be stored in a separate plugin repository or repositories. Plugin repositories can be housed and managed centrally for the entire PLN, locally, or a combination of the two. The MetaArchive Cooperative, for example, has deployed its plugin repository in a cloud computing environment, which allows for centralized location of the plugins, but provides a decentralized shared location for access and submission. The plugin repository or repositories should be placed under some kind of version control. LOCKSS makes use of a Concurrent Versions System (CVS), and the MetaArchive Cooperative makes use of Subversion. This allows for on-going changes to be documented, and if necessary to revert to earlier configurations.

Plugins must also be tested on a test cache or a test network to ensure that their crawl configuration successfully ingests the collection. This test network will be discussed in greater detail in Chapter 6: Content Ingest, Monitoring, and Recovery.

Once created and tested, plugins are packaged into files called JARs (an acronym for Java™ ARchive), which are signed and stored in a final plugin repository that is accessed by LOCKSS before it initiates an ingest on a cache. The LOCKSS daemon initiates ingest by accessing the title database to locate the proper plugin, its parameters, and the base URL from which to begin collecting or re-crawling content.

CONTENT MANAGEMENT

A Case Study in Preparing Content for LOCKSS Preservation

In the process of accumulating digital collections it is normal for directory structures, naming conventions, and metadata forms to become highly idiosyncratic, outmoded, and a hindrance to preservation readiness. When the focus turns to digital preservation readiness, then institutions become aware of the long-term detrimental effects of ad hoc preparedness.

For example, in 2008, the MetaArchive Cooperative and the Networked Digital Library of Theses and Dissertations (NDLTD) formed an alliance to examine the practical issues involved in a collaborative replication strategy for the digital preservation of ETDs. Shared below, the findings from that effort help to clarify the need for digital preservation readiness.

Preserving Restricted and Withheld ETDs

As previously mentioned in the section titled “Accessibility” above, to add a digital object to a PLN, it must be web-accessible (i.e., available via the HTTP or HTTPS protocol). PLN ingest requires a standard HTML permission-to-preserve statement on the host containing the ETD directory on its manifest page (see section titled “Manifest Pages” above).²⁴ While the manifest page is human readable, it is used entirely for programmatic ingest by the preservation network contributors. When access restrictions have been placed on some ETDs (for example, host university-only access), a list of specific content contributors’ IP addresses can be added to the web server’s firewall configuration to allow ingesting by only the specific network caches.

Structuring New ETD Collections for Ingest and Recovery

Organizing a contributing institution’s ETDs most effectively for preservation ingesting relies upon the creation of a methodical structure, such as a directory for each year’s ETDs. For larger institutions that approve hundreds of ETDs each year, annual directories should be further subdivided into logical units such as semesters or months. Smaller institutions that approve 100 or fewer ETDs per year will not benefit from creating these subdirectories.

While 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. Adopt a common, easy-to-decipher naming convention; for example, year/month 2008/01, 2008/02, etc. Remember, however, that these units are for programmatic ingest and not for human browsing.

When every contributing ETD member follows the same conventions for directory structure and file naming, each collection can be handled by a single plugin with different base URL and year parameters (see the section titled “Plugins” above). This consistency enables the network to provide members with effective but generic plugins. Otherwise, each institution must generate a plugin specific to its structure. The goal is thus to standardize naming conventions for files and directory structures from the beginning of any project. This will require analyzing the ways that the collection may grow over time, scoping numbering systems that can be parsed automatically, and developing of directory structures that can be easily traversed by subsequent ingesting systems. Data structures should also ideally be aligned with item-level metadata (see the section title “Metadata” above).

Successful ingest will depend on the content contributor’s ability to structure content into manageable AUs (see the section titled “Defining Archival Units” above). Each contributing institution needs to consider how the preservation copies stored in the network might be used to repopulate its existing archival structure in the future. A benefit of LOCKSS requires institutions to address preservation readiness at the point of ingest. For the network to easily ingest the content, a contributing institution is advised to have that content organized and well structured. Contributing institutions should remember that whatever work they have done to export the files and folders out of its repository system will need to be done in reverse in order to use them to repopulate that system.

URNs for ETDs

As students submit their ETDs, the files should be assigned a unique directory identifier with a Uniform Resource Name (URN). For example, an ETD submission that began at 5:57:13 on March 7, 2001 might become `etd-03072001-175713/`, based upon the date and time of submission. In this example, the ETD submission began at 5:57:13 on March 7, 2001. After an ETD is approved, the file(s) become part of the local collection. If an ETD that requires temporary embargo is approved, the upper directory structure

would be somewhat different, but the URN would be structured in the same predictable way. For example, an effective plugin would direct ingest from a given URL to find all the 2001 ETDs with instructions to:

1. Ignore the four numerals and '-' immediately following "etd" (i.e., -0307),
2. Recognize the year (in this case, 2001); and
3. Ignore the remaining characters.

This ETD, whether it has one file or many, would be placed in the school's 2001 AU. With this process, each year's-worth of ETDs is readily identifiable from each URN and can be divided into AUs by year on the preservation network caches without any data wrangling.

Triage for Legacy Collections

But what about collections that have been subjected to multiple repository conventions and those that straddle the gap between digitized and born-digital ETDs? Using Virginia Tech (VT) as a case study, the following demonstrates remediation approaches for entrenched ETD collections.

ETDs approved at VT before 2000 were named using a variety of URN conventions, such as /etd-454016449701231/ and /etd-030999-145545/. These URNs were not clearly structured or consistent though etd-030999-145545/ was probably submitted in 1999 and etd-454016449701231/ was most likely submitted in 1997. The solution was to establish a virtual collection with one AU for all pre-2000 ETDs. Plugin instructions were set to find all ETDs that did not fit the post-1999 URN convention. The complexity of this largely static collection is ultimately best served by plugin rules that exclude anything that matches the post-1999 format and places it into the PLN in an "Early VT ETD Collection."

Because digitized (as opposed to born-digital) bound theses and dissertations (BTDs) often follow the establishment of an ETD initiative, there exists a welcome opportunity to learn from earlier experiences. Scanned theses and dissertations can follow the URN naming convention based upon their digitization dates, rather than the dates on which they were originally approved. For example, a dissertation that was completed in 1994, but scanned Oct. 2, 2007 at 2:48:46 would be ingested with the existing plugin and

preserved in an assigned AU with the born-digital ETDs submitted in 2007. This method allows the static collection to remain unchanged. This system works for preservation purposes; however, it may need further consideration for rebuilding a public ETD database or collection from the preservation cache because works will likely be difficult to programmatically identify when reestablishing an annual grouping based on year of completion/approval.

It would not be very complicated on a conceptual level to programmatically generate URNs for BTDs based on their completion date, as this information likely exists in the contributing institution's MARC bibliographic records. BTDs are often assigned Library of Congress call numbers that also include dates. For example, the Limoges dissertation has the call number LD5655.V856 1994.L556. These call numbers are constructed as follows: Institution number--LD5655, thesis/dissertation number--V856, year--1994, Cutter number--L556.

In addition to file naming, batch processing involves pulling the physical items from possibly multiple locations (e.g., main library and remote storage). The process of arranging and maintaining their order, accurately deriving the file names from the MARC records, and linking them to the appropriate BTD files would become overly cumbersome and inefficient.

Final Remarks

Some PLNs, like the MetaArchive Cooperative, separate the function of the preservation caches from locally accessible collections. If it becomes necessary to rebuild a database of digital theses and dissertations (both digitized and born-digital) at the originating institution, the restoration of access, arrangement, and/or display of ETDs and BTDs is largely external to the purpose of the PLN. The goal of this case study has been to highlight the importance of organizing digital collections in ways that optimize both ingesting and repopulating a contributor's collections in the event of catastrophic loss – acknowledging that in every extant preservation solution there are going to be some trade-offs. The benefit of the LOCKSS solution is that it not only requires a contributor to become proactive in their digital preservation readiness, but also provides them with a sufficient amount of flexibility to carry out the preservation in ways that are best suited to their content and priorities.

CONCLUSION

In order to ensuring the successful ingest of content into a PLN, a content contributor must pay careful attention to its content's structure prior to its submission for harvest. This chapter has stressed the importance of pursuing preservation readiness before pursuing preservation itself. In the next chapter, the benefits of this expended effort will become more apparent as the process of collection ingest, monitoring, and recovery is covered in greater detail.

ENDNOTES

1. Library and Archives Canada's Digital Repository Services and Standards Office, *The Unified Digital Formats Registry*, 2009 <http://www.udfr.org/> (last accessed 12-14-2009).
2. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections," *Digital Preservation*, 2007 <http://www.digitalpreservation.gov/formats/> (last accessed 12-14-2009).
3. Rog, Judith; van Wijk, Caroline, "Evaluating File Formats for Long-term Preservation," *Koninklijke Bibliotheek*, 2008, 2. http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf (last accessed 12-14-2009).
4. Digital Library Foundation/OCLC Registry of Digital Masters Working Group, "Registry of Digital Masters Record Creation Guidelines," *The Digital Library Federation*, 2007 <http://www.diglib.org/collections/reg/DigRegGuide200705.htm> (last accessed 12-14-2009).
5. MetaArchive Cooperative; Networked Digital Library of Theses and Dissertations, "Electronic Theses and Dissertations Preservation Survey," *MetaArchive Cooperative*, 2008 http://www.metaarchive.org/public/resources/ndiipp_docs/NDIIPP_Market_Analysis.pdf (last accessed 12-14-2009).
6. CLOCKSS (Controlled LOCKSS): <http://www.clockss.org/clockss/Home> (last accessed 12-14-2009).
7. MetaArchive Cooperative: <http://www.metaarchive.org> (last accessed 12-14-2009).
8. Alabama Digital Preservation Network (ADPNet): <http://www.adpn.org/> (last accessed 12-14-2009).
9. LOCKSS (Lots of Copies Keep Stuff Safe): <http://www.lockss.org/> (last accessed 12-14-2009).

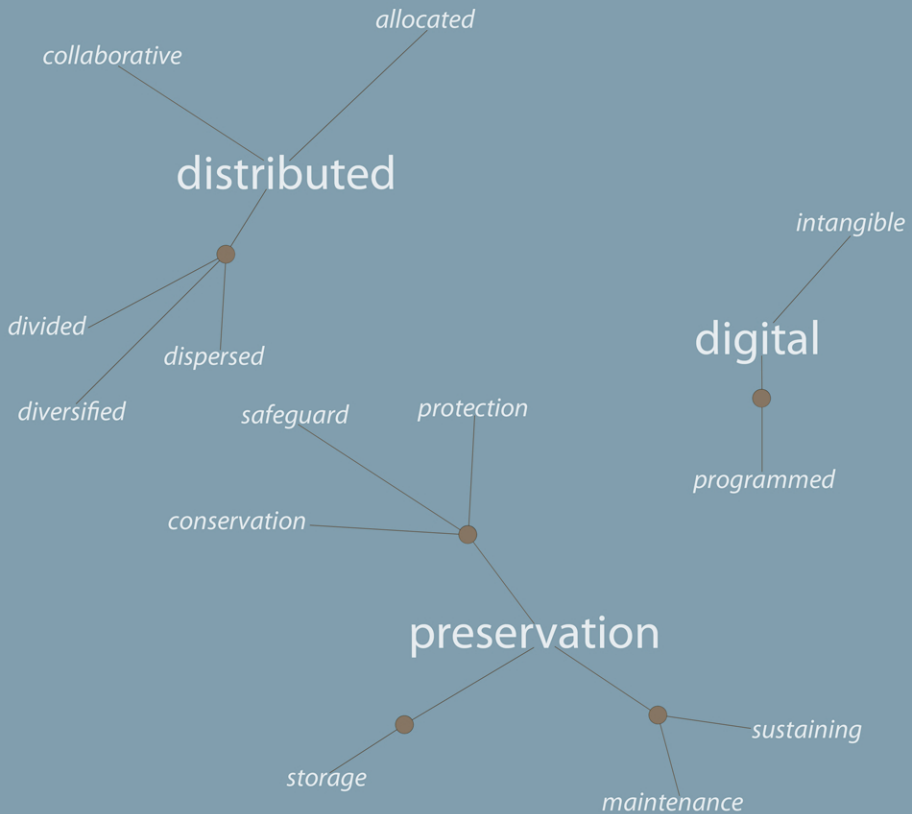
10. The Section 108 Study Group has been charged with updating for the digital world the Copyright Act's balance between the rights of creators and copyright owners and the needs of libraries and archives. This site has both an Executive Summary and the Full Report of the Study Group: <http://www.section108.gov/> (last accessed 12-14-2009).
11. The Membership Agreement of the MetaArchive Cooperative is a formal legal agreement for members to represent and warrant that, to the best of their knowledge, they are not contributing content to the preservation network that would infringe the rights of others: http://www.metaarchive.org/sites/default/files/Membership_Agreement_2010.pdf (last accessed 12-14-2009).
12. Lawrence, Gregory W.; et. al., "Risk Management of Digital Information: A File Format Investigation," *Council on Library and Information Resources*, 2000 <http://www.clir.org/pubs/reports/pub93/contents.html> (last accessed 12-14-2009).
13. Digital Repository Audit Method Based on Risk Assessment (DRAMBORA): <http://www.repositoryaudit.eu/> (last accessed 12-14-2009).
14. The intent of the CDP Dublin Core Metadata Best Practices (CDPDCMBP) is to provide guidelines for creating metadata records for digitized cultural heritage resources that are either born digital or have been reformatted from an existing physical resource. This document uses the Dublin Core element set as defined by the Dublin Core Metadata Initiative (DCMI): <http://www.bcr.org/dps/cdp/best/dublin-core-bp.pdf> (last accessed 12-14-2009).
15. The Dublin Core Collections Application Profile specifies how to construct a collection level description. It provides a means of creating simple descriptions of collections suitable for a broad range of collections, as well as simple descriptions of catalogues and indexes. Aggregations of physical or digital resources (collections) and aggregations of the metadata that describe them (catalogues and indices) can be described with similar properties: <http://dublincore.org/groups/collections/collection-application-profile/2007-03-09/> (last accessed 12-14-2009).
16. UKOLN Research Support Libraries Programme (RSLP) Collection Description Schema proposes a collection description schema, a structured set of metadata attributes, for describing collections within the RSLP: <http://www.ukoln.ac.uk/metadata/rspl/schema/> (last accessed 12-14-2009).
17. The IMLS DCC Collection Description Metadata Schema is based on the UKOLN RSLP Collection Description Metadata Schema and the Dublin Core Collection Description Application Profile. The IMLS DCC project has adapted these schemas to reflect the particular

nature and needs of the included projects. It is meant to describe digital collections created through IMLS Grant projects and does not describe in detail the projects themselves. This metadata schema forms the basis of the IMLS DCC Collection Registry: http://imlsdcc.grainger.uiuc.edu/CDschema_elements.asp (last accessed 12-14-2009).

18. PREMIS Preservation Metadata: Implementation Strategies is an initiative aimed at defining a core set of semantic units that repositories should know in order to perform their preservation functions: <http://www.loc.gov/standards/premis/> (last accessed 12-14-2009).
19. The MetaArchive Collection-Level Conspectus Metadata Specification charts and defines in detail the thirty-five elements that describe each collection: http://metaarchive.org/sites/default/files/conspectus_md_2005.html (last accessed 12-14-2009).
20. The following is a link to the LOCKSS plugin generation tool, which provides a user interface for creating and testing a plugin: http://www.lockss.org/lockss/Plugin_Tool (last accessed 12-14-2009).
21. The following is a link to brief overview of how to use the LOCKSS Plugin Tool: http://www.lockss.org/lockss/Plugin_Tool_Tutorial (last accessed 12-14-2009).
22. <http://metaarchive.org/metawiki/index.php?title=PluginStandards> (last accessed 12-14-2009).
23. The following is a link to a mini LOCKSS Plugin Tutorial that was developed by Virginia Tech graduate student Kamini Santhanagopalan: <http://scholar.lib.vt.edu/lockss/introduction.htm> (last accessed 12-14-2009).
24. For an example of a manifest page in the context of the discussion on "Preparing Content for LOCKSS Preservation" see this link to the Virginia Tech ETDs LOCKSS Manifest Page: <http://scholar.lib.vt.edu/theses/lockss/manifest.html> (last accessed 12-14-2009).

A Guide to Distributed Digital Preservation

Edited by Katherine Skinner
and Matt Schultz






Copyright © 2010

This collection is covered by the following Creative Commons License:

Attribution-NonCommercial-NoDerivs 3.0 License

You are free to copy, distribute, and display this work under the following conditions:

	Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). Specifically, you must state that the work was originally published in <i>A Guide to Distributed Digital Preservation</i> (2010), Katherine Skinner and Matt Schultz, eds., and you must attribute the individual author(s).
	Noncommercial. You may not use this work for commercial purposes.
	No Derivative Works. You may not alter, transform, or build upon this work.

For any reuse or distribution, you must make clear to others the license terms of this work.

Any of these conditions can be waived if you get permission from the copyright holder.

Nothing in this license impairs or restricts the author's moral rights.

The above is a summary of the full license, which is available at the following URL:

<http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode>

Publication and Cataloging Information:

ISBN: 978-0-9826653-0-5

Editors: Katherine Skinner
Matt Schultz

Copyeditor: Susan Wells Parham

Publisher: Educopia Institute
Atlanta, GA 30309

TABLE OF CONTENTS

Acknowledgements	vii
Road Map	ix
Chapter 1: Preserving Our Collections, Preserving Our Missions	1
Chapter 2: DDP Architecture	11
Chapter 3: Technical Considerations for PLNs.....	27
Chapter 4: Organizational Considerations.....	37
Chapter 5: Content Selection, Preparation, and Management.....	49
Chapter 6: Content Ingest, Monitoring, and Recovery for PLNs.....	73
Chapter 7: Cache and Network Administration for PLNs.....	85
Chapter 8: Copyright Practice in the DDP: Practice makes Perfect.....	99
Appendix A: Private LOCKSS Networks	113
Glossary of Terms	127
Author Biographies	133