
MetaArchive of Southern Digital Culture

LC NDIIIP Proposal
FINAL REVISED VERSION

This document includes all descriptive narrative appendices from the MetaArchive
NDIIP project agreement with the Library of Congress.

A.2 Project Description

A.2.1 With support from the Library of Congress, the partner institutions of this project envision a three-year process to develop a cooperative for the preservation of at-risk digital content with a particular content focus: the culture and history of the American South. The project group members will jointly develop: 1) a prioritized conspectus of at-risk digital content in this subject domain held at the partner sites, 2) a harvested body of the most critical content at the partner sites to be preserved, 3) a cooperative agreement for ongoing collaboration, and 4) a distributed preservation network infrastructure based on the LOCKSS software. The proposed work plan for this project builds on relationships and workflows developed during previous projects of the MetaScholar Initiative and other collaborating consortia.

A.2.2 Approach to Digital Preservation

A.2.2.1 The MetaArchive approach to long term digital preservation of at-risk content will initiate coordinated efforts by a decentralized group of peer institutions to mutually identify, preserve, and archive information. While it is questionable whether any centralized scheme for preservation can succeed over the long haul, decentralized mechanisms for mobilizing group efforts between cooperating institutions appear to hold promise for effective long-term models for preservation. The success of the Internet itself is an example of why de-centered peer-to-peer approaches are often more robust and effective than hierarchical organizational architectures.

A.2.2.2 The MetaArchive project proposes an initial investigation of a cooperative model of peer institutions partnering for digital preservation purposes. This cooperative will build a network of collaborating institutions acting to preserve a corpus of critical cultural heritage content available only in digital forms. This preservation network will use the LOCKSS software, a major new system for distributed preservation of digital content. Collaborative partnership models based on this approach may be the best hope for successful digital preservation of our cultural heritage.

A.2.2.3 This project is a first step in advancing the new practice of digital preservation. In the long run, preservation of digital content will require much more than simply identifying and archiving content. Preservation efforts will require automated mechanisms for format migration, metadata validation, sophisticated rights management, and many more functions. But before these more advanced steps are taken, a foundation of practices is needed for distributed archiving and content conspectus building. Once this preservation network is established, we envision future projects to investigate and advance these additional areas of digital preservation.

A.2.3 Key Features of a Secure MetaArchive

A.2.3.1 This project seeks to cooperatively establish a secure digital archive of digital archives, or MetaArchive. Distinctive features of this project include:

1. **Distributed Preservation.** Most effective digital preservation efforts in practice succeed through some strategy for distributing copies of content in secure, distributed locations over time. By basing the approach to digital preservation on a leading preservation software package for distributed digital replication (LOCKSS), this preservation network will establish from the beginning a distributed means of replicated archives.
2. **Flexible Organizational Model.** This project will develop a simple and flexible cooperative agreement as a model for other institutions seeking to cooperate for purposes of digital preservation. This agreement will entail minimal overhead, enlist straightforward mechanisms for collaboration, and be widely applicable to many sorts of institutions. By creating a cooperative model for digital preservation that is intentionally designed for broad applicability to many other settings, the cooperative seeks to not only create an effective preservation network for one body of digital content, but enable the creation of many others for this important purpose.
3. **Content Selection.** The cooperative model of this project includes the formation of a group of content experts from the partner libraries who will guide the development of the subject domain conspectus for at-risk digital content. This group will be composed of librarians, archivists, and specialists in the technical issues of digital preservation. Prioritization of at-risk content must be conducted by such a team, with diverse backgrounds and experiences brought to bear on the difficult question of selecting and prioritizing the digital content for preservation. The team will evaluate content at the partner sites in terms of its importance for cultural heritage, degree to which the content is at-risk, and preservation considerations (including format and planning for text and image migration during the initial conspectus).
4. **Migrating Archives.** In order to be preserved, digital content must be maintained in a manner that facilitates migrating it over time. Although digital content migration is still poorly understood and findings are preliminary, some factors clearly lend themselves to long term migration. One is ensuring that the software used for such migration is open source, so that the software itself can be preserved and evolved. Second, metadata concerning the archived content must be carefully maintained. Finally, strategies for preserving data through its storage in migratable formats and data structures. These measures will lay the basic groundwork for subsequent migration efforts.
5. **Dark Archiving.** Many preservation efforts conflate maximizing short-term access (high availability) with long-term access (preservation). High availability entails strategies for ensuring that content is available constantly for highly visible public downloads, on 24x7 systems with tremendous attention devoted to “up time.” The cooperative will forego the expense and effort of high availability measures, and focus instead on long-term preservation issues. One or more nodes in our preservation network may be down at any particular time, but the content collectively maintained in

the preservation network will always be secure. Content held in the preservation network will be discoverable by means of standard metadata dissemination mechanisms (the Open Archives Initiative Protocol for Metadata Harvesting), but it will not necessarily be immediately downloadable on demand by all comers external to the network. The cooperative will develop a process for retrieving items from the preservation network, while carefully checking and tracking released information.

6. **Relatively Low Cost.** This approach to digital preservation is intentionally designed to require minimal expenditures by collaborating groups of medium sized institutions. Building on the LOCKSS approach of low-cost, low barriers to adoption, the proposed preservation network should be a model that can be easily implemented by many ad hoc groups of collaborating institutions. The cooperative will develop a freely available, open source adaptation of the LOCKSS software (which is free and open source software). This software will run on inexpensive computers, and will require only a relatively modest degree of systems administration for ongoing maintenance. This assertion reflects the knowledge that the LOCKSS software upon which the proposed system is based has these characteristics.
7. **Self-Sustaining Incentives.** Although the preservation network created in this project will serve to preserve content of benefit to the common good of the American public, it will also have several built-in institutional incentives to ensure its independent sustainability over time. Beyond the low-cost features described previously, the preservation network will be a desirable service for participating institutions to continue maintaining over time, as it will provide institutions with a cooperative mechanism for preserving content of not only shared value, but also particular (critical) value to the individual institutions. The preservation network accomplishes this by providing each institution with a modest quota for submitted content selected solely by the institution, yet securely maintained jointly by all institutions participating in the network. This ensures that the network will always provide the participating institutions with a capability that they fundamentally lack individually.
8. **Simple Preservation Exchange Mechanisms with the Library of Congress.** The distributed and automated approach of this project to digital preservation simplifies the mechanism for sharing the resulting archive of digital content with the Library of Congress. The straightforward process for replicating the content of the MetaArchive is to install and activate another standardized node of the preservation network at the Library of Congress. Validated, ongoing replication of the exact content of the MetaArchive at the Library of Congress is ensured by the design of the system.

A.2.4 Categories of At-Risk Digital Content under Consideration

A.2.4.1 There are many varieties of digital content concerning the culture and history of the American South that are critically important to understanding our national cultural heritage. The Library of Congress has devoted collection development efforts to many of these topics, including such broad subjects as the Civil War, slave narratives, the Civil Rights Movement, regional accounts from Southern states, Southern music, Southern handicrafts, and Southern church histories. All of these topical areas have seen a tremendous growth in recent years of digital materials which are unique and at-risk. Although many of these materials have been developed starting with digitized forms of analog materials, they all incorporate uniquely valuable digital aspects for which no analog equivalent can exist.

A.2.4.2 Examples of collections under consideration for preservation include the following:

- a. *Digital Masters* of scanned images of brittle or decaying analog originals are usually not considered at-risk, and yet David Seaman (Director of the Digital Library Federation) has pointed out that preservation efforts for such materials are often in fact inadequate, as what actually receives attention in digitization projects is preservation of derivative digital forms that comprise the working online systems in digital libraries. Digital masters are most often stored on CD-ROMs in closet spaces with no systematic efforts to validate the integrity of the data over time.
- b. *Research Databases and other digital content* created by scholars working in the field are increasingly a rich source of primary source information about our cultural heritage. Ethnographic investigations now routinely produce databases of images, digitally recorded interviews, and many other categories of invaluable information. However, in most cases, scholars do not systematically preserve this information.
- c. *Website Exhibitions* that provide digital perspectives on cultural issues. Such digital exhibitions can bring static artifacts and dead history to life through animations, hyperlinks, integrated commentary, and many other digital features. Yet, such digital exhibitions are often considered ancillary to a physical exhibition, and frequently do not receive effective preservation attention.

A.2.4.3 Additional categories of at-risk digital content include local institutional databases, ad hoc websites, and others. All of these varieties of digital content illuminate our cultural history and would be much more likely to survive for posterity if systematic preservation efforts were undertaken on their behalf.

A.2.5 Roles and Allocation of Responsibilities

A.2.5.1 For this collaborative peer group approach to work well, categories of roles and responsibilities among the members of the collaboration must be clearly understood from the beginning. The following are key entities and roles in the project.

A.2.5.2 Individual Roles:

- a. *Principal Investigators* will be responsible for leading the project, acting as project managers of efforts at their respective sites, and forming the membership of the Steering Committee (see below). Each Preservation Site will include at least one Principal Investigator. Two individuals will jointly share the PI role at FSU, and will trade off participation on the Steering Committee. There will therefore be a total of seven Co-Principal Investigators.
- b. *Content Consultants* will be individuals who are also knowledgeable concerning relevant digital collections at the partner institutions. They may be archivists or curators responsible for major collections, or librarians with a subject background in some broad area of Southern studies. They may be digitization specialists working specifically with Southern materials. They may be programmatic leaders of research centers or regional consortium services focusing on Southern cultural heritage topics. They will all have experience with the digital collections held at the partner sites.
- c. *Technical Consultants* will be individuals with a background that prepares them for advanced work on digital preservation activities of the project. Members of this group may be digital preservation project experts, computer scientists who have experience with the research aspects of digital preservation, systems librarians who have significant experience with digitization projects and the long-term maintenance of the content in such systems, or other advanced technical professionals who have worked intensively on digital preservation technologies, especially in the library or archival setting.
- d. *Project Staff* will be individuals who will work on this project in various capacities, whether content-related or technical in nature. Their salaries may be funded either by grant allocations as in-kind cost match by project institutions. Project staff may serve on the Content and Preservation Committees.

A.2.5.3 Committee Roles

- a. *The Steering Committee* will be responsible for overall project coordination, meeting deadlines, and project communication and reporting efforts. The Steering Committee will be comprised of the Principal Investigators acting as a group.
- b. *The Content Committee* will be responsible for organizing, developing, and conducting content identification/selection practices, as well as prioritization of content, and organizational aspects of partnership building. The Content Committee will include all Principal Investigators as well as some additional advisory staff.
- c. *The Preservation Committee* will be responsible for developing technical and evaluative means for digital preservation, practices for content acquisition, metadata tracking, mechanisms for content retention/transfer, and technical aspects of partnership building. Such strategies will include those for regulating text and image structures to best ensure continued migratability. The Preservation Committee will include all Technical Consultants, some project staff, and some Principal Investigators.

A.2.5.4 Institutional and Consortial Roles

- a. *Preservation Sites* will be the primary entities responsible for the ongoing activity of preserving digital content. At a minimum, a preservation site must include a node server of the Preservation Network and either a principal investigator or the AOTR. These node servers will each use a modified form of the LOCKSS software. The cooperative envisions seven Preservation Sites as part of this project (including the Library of Congress). The Preservation Sites collectively comprise the Preservation Network. Auburn and Louisville will serve solely as preservation sites.
- b. *Development Sites* will be responsible for all of the functions of a Preservation Site, as well as technical development of the computer systems that enable the Preservation Network. The major responsibilities of this development activity will include adaptation of the LOCKSS software for preservation of relevant digital content, design and implementation of system features dealing with security, content validation and integrity checking, and metadata tracking. Emory, GA tech, VA Tech, and FSU will serve as development sites.
- c. *The Preservation Network* (also termed the *Cooperative*) is composed of all Preservation Sites that work together to preserve at-risk digital content. This Preservation Network functions as a MetaArchive, or archive of archives of such content; hence the name of the project.

A.4 Key Personnel

A.4.1 There are a large number of persons who will work on this project. However, the only individual who should be considered irreplaceable without prior approval from the Library of Congress is the Project Manager, Martin Halbert (Director for Library Systems, Emory University). There will undoubtedly be staff turnover and concomitant replacements that will occur among the other individuals mentioned in the working committees below in the course of this three year project; such changes will be reported in the project reports as they occur. The following are the currently envisioned committee and institutional assignments.

A.4.2. Steering Committee. *Martin Halbert* (Director for Library Systems, Emory) will serve as the lead Principal Investigator and will chair the Steering Committee, as well as serving on the Preservation and Content Committees. Emory will serve as the lead institution of the project, and a project Development Site. Halbert has been Executive Director of the MetaScholar Initiative since its inception. *Tyler Walters* (Assoc. Director for Digital and Technical Services, GA Tech) will serve on both the Content and Preservation Committees. GA Tech will be a project Development Site. *Robert McDonald* (Asst. Dir. for Technology Services, FSU Libraries) and *Chuck Thomas* (Head of the FSU Digital Library Center, Florida State University) will jointly serve on the Preservation Committee. Florida State will be a project Development Site. *Delinda Buie* (Curator of Rare Books, University of Louisville) will serve on the Content Committee. Louisville will be a project Preservation Site. *Gail McMillan* (Director of Digital Library and Archives, VA Tech) will serve on both the Content and Preservation Committees. Virginia Tech will be a project Development Site. *Beth Nicol* (Information Technology Master Specialist, Auburn) will serve on both the Content and Preservation Committees. Auburn will be a project Preservation Site.

A.4.3 Content Committee. All Steering Committee members will serve on the Content Committee. Additionally, the following individuals will contribute and advise, although not necessarily in person during project meetings. *Dr. Pablo Davis* (SAHC Director) for Virginia Tech collections. *Naomi Nelson* (Curator for Southern Collections, Emory) will work on conspectus development. Catherine Jannik (Digital Initiatives Manager, GA Tech) has a scholarly background in the history of the South, and will work on conspectus development, as well as cooperative agreement analysis. *Dwayne Butler* (Assoc. Prof., Endowed Chair, Scholarly Communication, Louisville) will analyze intellectual property issues for the conspectus and harvesting efforts. *Aaron Trehub* (Director of Library Technology, Auburn) will work on conspectus development, with particular focus on Alabama resources. *Henry McCurley* (Chair, Cataloging Dept., Auburn), will contribute metadata expertise to conspectus development activities.

A.4.4 Preservation Committee. All Steering Committee members may serve on the Preservation Committee. Technical Consultants: *Thomas Robertson* (LOCKSS Project) is a key project leader of the international LOCKSS project. He has intensively studied the question of how a distributed network of systems can work together for long term preservation of content. *Dr. Edward Fox* (Digital Library Research Lab and CITIDEL Director, Virginia Tech) has

served as the PI on dozens of digital library research projects and is an international authority on digital library technologies. Additional Project Staff who may advise, although not necessarily in person during project meetings include: *Lars Meyer* (Head of Preservation, Emory) will analyze digital preservation issues and provide in depth expertise in strategies for *both* preservation through content structures and archiving. *Susan Coleman* (Head, Systems Dept., GA Tech) will assist in developing the cooperative agreement, and assist in planning and implementing the technological aspects of GA Tech's participation in the preservation network. *Larry Hansard* (Programming & Networking Manager, GA Tech) will participate in LOCKSS software adaptation and management, server administration, and graduate assistant supervision. *Susan Parham* (Digital Initiatives Tech. Ops. Coord., GA Tech) will work on developing the technical routines that the cooperative will follow to manage the stored digital objects, including metadata needs. *Curtis Carr* (Dir. Univ. Lib. Systems Dept., VA Tech) will be responsible for systems administration, local software testing, and continuing resources for local and regional support. *Jason Knight* (Web Development Librarian, FSU) will work on preservation network development and deployment. *Weiling Liu* (Dir. Lib. Tech, Louisville) will work on preservation network deployment and synchronization issues.

A.5 Approved Phases

A.5.1 The project will take place over a period of three years and will be oriented around four phases corresponding to the four major deliverables.

A.5.2 PHASE I: Content Identification and Selection

A.5.2.1 The primary deliverable of this project phase is the *Content Conspectus*. During this project phase the Content Committee will be convened and conduct its work of identifying and prioritizing at-risk digital content.

A.5.2.2 Work of the Content Committee: A database of content sites under consideration by the committee will be completed during this phase of work to create the initial conspectus. Issues for the Content Committee to resolve and document during this project phase include a process for selection of sites to be preserved, MIME media types to be included and excluded from consideration, and additional details concerning intellectual property, access restrictions, and long-term preservation/migration of content. This work will be coordinated with the Preservation Committee by means of several of the principal investigators who will serve on both committees. All-project meetings of both committees will serve as milestones associated with key outcomes.

A.5.2.3 Aims of the Conspectus: The initial conspectus will guide the initial content harvest during the second phase of the project. The initial conspectus should seek to identify the most prominent and critical digital content sites in need of preservation, but will not attempt to identify all sites exhaustively. The initial conspectus will be enhanced and revised during the second year of the project through further work to discover less-prominent but still important at-risk sites, and to establish guidelines for future development.

A.5.3 PHASE II: Content Acquisition

A.5.3.1 The primary deliverable of this project phase is the *Content Harvest*. During this project phase the Preservation Committee will work to develop and conduct the technical mechanism for gathering and preserving the content harvest. This phase will be conducted in parallel with the preparation of the content conspectus.

A.5.3.2 Capture Mechanism: The LOCKSS software will be extended for use in preserving several varieties of digital content decided upon by the Content and Preservation Committees. The LOCKSS software has built in harvesting modules that were designed to harvest any MIME media type. These harvesting modules will be activated, tested, and used as the capture mechanism. The entire preservation network need not be operational for the initial harvest to take place. Depending on the progress made by the various sites, the harvest may be undertaken by a single preservation node or several working in parallel, as the process is modular, and portions of the harvest can be easily assembled to form the entire content harvest.

A.5.3.3 Content Authentication Approach: LOCKSS incorporates the RSA MD5 data

W310775.6

validation algorithm to verify the integrity of harvested data subsequent to harvest and replication. In addition to this mechanism, validation scripts will be considered for application to the harvested content. Testing and assessment of the harvested content will be undertaken to identify any problems in the content metadata, synchronization of the network nodes, and security considerations.

A.5.4 PHASE III: Partnership Building

A.5.4.1 The primary deliverable of this project phase is the *Cooperative Agreement*. During this project phase the Steering Committee will analyze and develop a model agreement describing ongoing roles and responsibilities of the members of the Preservation Network Cooperative. This agreement will describe how this group of institutions can work together in the future to preserve the body of harvested digital content resulting from the project.

A.5.4.2 Process for Drafting the Agreement: Analysis of the various relevant aspects of the cooperative will be undertaken during this period. This analysis will study how the cooperative is working so far, problem areas, unforeseen issues, and topics for further scrutiny. The analysis will inform the outline of the agreement and the drafting process. The agreement will be finalized by the end of the project.

A.5.4.3 Broader Applicability: The Steering Committee will seek to ensure that the cooperative agreement not only models the functioning of this particular group of institutions for this particular purpose, but also models how small consortia and other groups of institutions, with relatively limited technical means, can work together to preserve digital content. The LOCKSS software is a freely available open source toolkit, and all of the software developed in this project also will be freely available open source software. The Steering Committee will engage in discussions with other consortial groups to seek feedback and perspective on how such an agreement can be made more broadly applicable.

A.5.5 PHASE IV: Content Retention/Transfer

A.5.5.1 The primary deliverable of this project phase is the *Preservation Network*. The modified version of the LOCKSS software will be installed on relatively low cost server clusters at each of the preservation sites. The preservation network will be activated and the content contained in the network will be synchronized, ensuring that the content harvest will be replicated and validated at each of the preservation sites.

A.5.5.2 Sharing the Content Harvest with the Library of Congress: Transmitting the Content Harvest to the Library of Congress is conceptually a straightforward process with the technology we will use. By installing a preservation network node at the Library of Congress, the entire content harvest plus all updates will automatically be transmitted and maintained in an up-to-date fashion. The Steering Committee will discuss this process with Library of Congress staff in the course of the project, as the issues of installing and maintaining such a preservation node become clearer. We will work with the staff of the Library of Congress to add the LC node to the network.

A.6 Approved Work Plans and Deliverables

A.6.1 Deliverables. The project will produce four major deliverables, each of which will comprise a project goal and be associated with one of the project phases. The deliverables are as follows:

A.6.1.1 Content Conspectus: This will be a comprehensive survey of at-risk digital content in the subject domain of Southern cultural heritage at the partner libraries, conducted by the members of the Content Committee. Subjects: Broad topical areas of digitized content relating to Southern culture and history will be examined in light of their relevance to the understanding of key issues in the culture and history of the American South, as well as to what extent they are complementary to major collecting areas of the Library of Congress. Materials: A variety of at-risk categories of materials will be studied. Formats and special consideration associated with each format will be studied in the course of developing the conspectus. Information Collated: The conspectus will take the form of an online database which only the members of the Content Committee can edit through authenticated access. The metadata in the conspectus may follow the new UKOLN RSLP Collection Description schema (see <http://www.ukoln.ac.uk/metadata/rsrp/schema>), which provides a framework for consistently and effectively describing collections. Following this emerging specification, the database will record data for each content site or collection concerning the three broad areas of A) collection information, B) location information, and C) agent information. All 46 attributes described in the current version of the RSLP schema will be considered for metadata fields in the conspectus database. Quality standards for conspectus metadata records will be agreed upon during the initial work of the Content Committee; specifically deciding which RSLP fields will be mandated in conspectus records. Beyond simple identification of collections, we anticipate that the conspectus database will minimally include information concerning the importance of specific collections and extent/size of the collection.

A.6.1.2 Content Harvest: During the course of the project, the cooperative will aggregate the content targeted in the conspectus. This content harvest will accumulate in stages, and the harvest process will not be considered complete until the end of the project. This content harvest will be shared with the Library of Congress through the establishment of a preservation network node at LoC. Modified LOCKSS Software and Ingest Mechanism: This material may be aggregated by several potential ingest mechanisms, but the primary means we anticipate investigating will be focused web crawls by the modified LOCKSS software that the project will develop. The LOCKSS software already includes a web content crawler that can easily be enhanced to load file systems of digital masters and other bodies of content. We will only seek to archive static representations of content, not complex behaviors associated with dynamic web systems (although this topic may be taken up in a future project).

A.6.1.3 Cooperative Agreement: The formation of this cooperative will include a detailed charter document for the MetaArchive preservation network, including an analysis and articulation of the collaborative paradigm that this project will investigate. This document will be designed in such a way that it can serve as a model and template for other groups of

institutions seeking to replicate this group's approach to distributed digital preservation.

Components of the Agreement: This charter will include sections addressing all 11 articles of collaboration mentioned in the NDIP solicitation under "Partnership Building" in the original solicitation. It will devote particular attention to describing the ongoing management structure recommended, membership criteria, roles and responsibilities, sustainability plan, and other details of particular concern to our organizational model.

A.6.1.4 Preservation Network: This will be the functioning network of modified LOCKSS server nodes that collectively will act to preserve digital content over time. The preservation network is intended to be a distributed, mutually administered archive of critical digital content that has internal mechanisms for data integrity checks and heavy security and fault-tolerance features. It will have metadata registries accessible to all contributing sites and the Library of Congress.

A.6.1.4.1 Security Characteristics of Preservation Network: We anticipate that the nodes of the preservation network will each be composed of two paired servers, termed the *gate* and the *vault*. The gate is a server functioning as a dedicated firewall for the preservation node. The gate server will deploy a hardened operating system, such as the security-enhanced Linux developed and distributed by the National Security Agency (see <http://www.nsa.gov/selinux/>), or the cryptographically hardened OpenBSD (see <http://www.openbsd.org/security.html>). Each vault server will run the modified LOCKSS software developed for the content harvest, and will also deploy a hardened operating system for in-depth protection of the preserved content. The vault server will not be connected directly to the Internet, but will instead be secured behind the gate server, and will interact with the Internet through only the gate server. Each gate server will respond only to specified network addresses and ports, and will have a minimal set of communication ports and services activated. This layered security provides an extremely secure architecture, and yet is relatively simple to maintain and monitor. Each vault server will have both mirrored system drives as well as a large 2 TB hard-drive array to store the preserved digital content.

A.6.1.4.2 Preservation Network Metadata Registries: We anticipate that the preservation network will have associated metadata registries providing information concerning the preserved content. These will take the form of a *collection registry* (this registry will simply be a subset of the conspectus, with records specifying which collections were actually harvested), and a *content registry* (recording metadata for each digital object aggregated during the harvesting process, including the RSA MD5 data validation hash entry for the modified LOCKSS software to track validity of the data). The collection registry will be made publicly available via a public OAI data provider system that will disseminate metadata concerning the collections upon demand through the OAI-PMH communication standard. The content registry will not be publicly available, since it contains security information, and will be accessible only to members of the cooperative. For security reasons, neither of the metadata registries will likely be maintained on servers comprising the preservation network, but instead will be separately kept on independent systems.

A.6.1.4.3 Preservation Network Architecture: The LOCKSS software is fundamentally a peer-to-peer (P2P) network architecture with functionality for data maintenance, integrity checking, and other features fundamental to digital preservation. All of the preservation nodes act as peers for the purpose of long term preservation of content (see figure below). Each node communicates with all other nodes of the preservation network and only these nodes. The only exception to this rule is when harvesting operations are underway, when one or more nodes of the network are ingesting content.

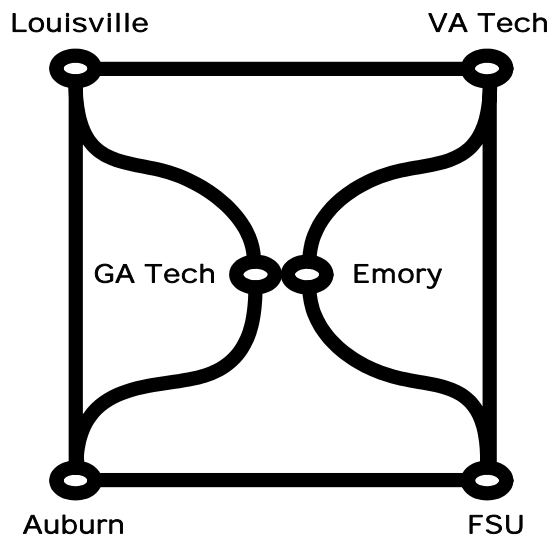
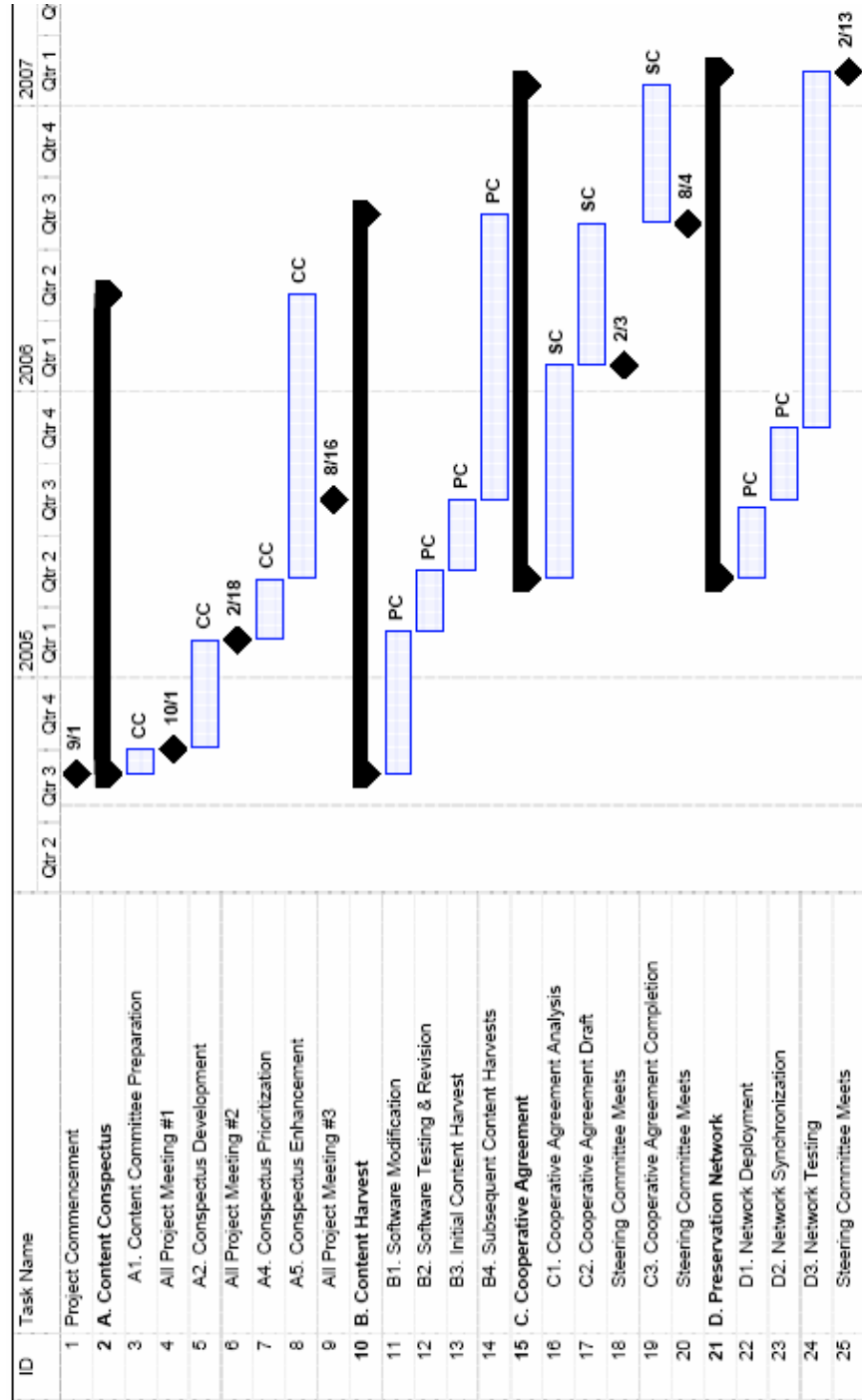


Figure 1: Preservation Network Diagram

A.6.1.4.4 Preservation Network Shared and Institutional Content Quotas: Institutional quotas are a feature of the preservation network designed to reinforce institutional support for ongoing maintenance of the network. The disk storage arrays attached to each vault server in the network should be capable of storing approximately 2.2 TB (or more, depending on equipment costs at time of purchase) of data in aggregate. A total of 2 TB (approximately 90%) of the network's 2.2 TB storage capacity will be dedicated to the shared content harvest that the cooperative will jointly identify and assemble. The remainder, some 200 GB (or roughly 10%) of the network's capacity will be allocated for preservation of critical content identified solely by the individual institutions making up the network. By allocating a quota of 32 GB of replicated, secure storage to each member of the cooperative for preservation of locally determined content, we hope to offer a clear incentive for members to continue in the cooperative in the future. By creating a mechanism for cooperative members to both contribute to the common good and individual interests, our hope is to strike an effective balance that will be sustainable over time.

A.6.1.4.5 Preservation Network Shared Custodial Responsibilities. It is important to reiterate that because of the design of the LOCKSS software, all preservation sites will serve as joint custodians of the content harvested. Should one of the sites withdraw from the cooperative, or simply become dysfunctional, no loss of data will occur, as the entire content harvest is reliably preserved and validated at all preservation Sites. In this way, the MetaArchive accomplishes what an archive of physical items cannot, specifically multiple replication at geographically distant and secure sites. The security of the MetaArchive is enhanced by the quantity of Preservation Sites taking part. It is worth noting that the six partners of the MetaArchive project are all also members of the Internet 2 community, meaning that loss recovery and other MetaArchive functions potentially can benefit from this association. Additional future members of the MetaArchive that join can serve to enhance the security of the overall preservation network. Because of the self-reinforcing approach to security instantiated in the LOCKSS software, the likelihood of data loss becomes statistically unlikely once six institutions are participating in the network. If additional preservation sites join, as we hope will be the case in the future, then the probability of data loss becomes increasingly unlikely.

A.6.2 Project Gant Chart



A.6.3 Milestones: Major project milestones will be marked by meetings of project personnel coming together to discuss issues and document results. Both milestones and work plans should be read in conjunction with the project schedule.

A.6.3.1 October 2004: Committee Formation Meeting. The first of three all-project meetings will be held to bring together the group of librarians, archivists, and technologists who will work on project activities. During this meeting the group will identify group and individual objectives for the Content and Preservation Committees, with mechanisms for regular status reports and progress checks.

A.6.3.2 February 2005: Harvest Preparation Meeting. The second of three all-project meetings will be held to finalize the initial conspectus that will guide the initial content harvest, review the results of software development, and to make final preparations for the first full content harvest.

A.6.3.3 August 2005: Preservation Network Inauguration Meeting. The final all-project meeting will be held to inaugurate the preservation network, refine the conspectus with additional sites, share initial analysis for the development of the cooperative agreement, and examine the content of the first harvest as it is disseminated across the network.

A.6.3.4 February 2006: Cooperative Agreement Draft Meeting. The Steering Committee will review the results of the analysis on cooperative models conducted, and reach consensus on the key elements of the Cooperative agreement. Additional meeting agenda items will include reviewing the results of the refined Conspectus and subsequent content harvests, as well as the early results from preservation network tests.

A.6.3.5 August 2006: Cooperative Agreement Finalization Meeting. The Steering Committee will work during this meeting to finalize the cooperative agreement. Results of preservation network tests will be reviewed. An outline of the final project report will be developed.

A.6.3.6 February 2007: Final Project Meeting. The Steering Committee will meet to complete the final project report.

A.6.4 Content Conspectus Workplan

Content Committee Preparation. Several steps will be taken to prepare for the work of the conspectus development, primarily planning and organization steps for the first all-project meeting. Most time consuming steps are coordination of travel arrangements, and detailed meeting agenda and other arrangements planning. Emory will take responsibility for organizing and hosting this first all-project meeting, although other meetings may be held at other partner sites.

All Project Meeting #1. *Meeting Activities:* Review project plan, conduct organizational formation sessions for the three committees, and undertake planning activities. Sessions of all project participants will be scheduled to coordinate all project work, and breakout sessions of the various committees will be held to advance their respective efforts. *Meeting Outcomes:* 1) Conspectus Plan, including detailed group and individual objectives and deadlines for conspectus development activities, 2) Communication Plan, including detailed mechanisms to keep all project personnel in regular communication via conference calls and email, with information tracked and archived, 3) Equipment Purchase Plan, including final configuration for preservation node equipment, and plans for ordering this equipment in preparation for next steps.

Conspectus Development. The Content Committee will work for 20 weeks to develop the first version of the conspectus. This work will be coordinated by means of the communication plan described above. Members of the Content Committee will identify content both singly and in group activities guided by the conspectus plan described above. The initial conspectus database and summary conspectus documentation should be completed in a timeframe such that it will be available in advance of the second all-project meeting for the initial prioritization decisions to be undertaken at this meeting.

All Project Meeting #2. The second all-project meeting will be focused on preparations for the initial content harvest. *Meeting Activities:* Review plan for the upcoming content harvest, demo and report on the software development activities to date, conduct planning sessions for the three committees, and undertake group project activities. Sessions of all project participants will be scheduled to coordinate all project work, and breakout sessions of the various committees will be held to advance their respective efforts. *Meeting Outcomes:* 1) Initial Conspectus, including priorities for the initial content harvest, 2) Initial Harvest Plan, including detailed timetable for harvest activities, reporting plan, and coordination with the network deployment plan, which will include the plan for synchronizing the content throughout the network, 3) Network Deployment Plan, including timetable for specific deployment steps at partner sites.

Conspectus Prioritization. This work will begin before the second all-project meeting, and will continue through 11 weeks. It is unknown at this point whether or not all content identified in the conspectus can be harvested in the initial harvest, whether because of space considerations, access restrictions, or other reasons. The prioritization of the conspectus will determine what the initial harvest specifically seeks to aggregate. The initial harvest should not be so large that it fills up the entire potential capacity of the preservation network, but rather should leave significant room for additional growth. Percentages and amounts are not specified here, in order to retain flexibility for the remainder of the project.

Conspectus Enhancement. The initial conspectus will continue to be developed through project month 24, through the identification of more content sites through the same mechanisms used for development of the initial conspectus.

All Project Meeting #3. The third all-project meeting will focus on assessment of the preservation network, which will have been deployed by this point. *Meeting Activities:* Review the deployment activities of the preservation network to date, enhance the conspectus, conduct planning sessions for the three committees, and undertake group project activities. Sessions of all project participants will be scheduled to coordinate all project work, and breakout sessions of the various committees will be held to advance their respective efforts. *Meeting Outcomes:* 1) Subsequent Harvest Plan, identifying the timeline for subsequent harvests, 2) Preservation Network Assessment Plan, which will identify specific assessment steps to be taken, 3) Cooperative Agreement Analysis Plan, describing what information will be gathered to inform the cooperative agreement development, and how this information will be analyzed.

A.6.5 Content Harvest Work Plan

Software Modification. A subset of the Preservation Committee consisting of technical staff at the development sites will work together on modification of the LOCKSS software as described in this proposal. Software modification and development will be coordinated by Emory University. 26 weeks have been allocated for this activity, which should be generous given the modest scale of the modifications anticipated. Also included in this period are related activities of firewall testing, and equipment specification development (which will inform the purchases planned at the first all-project meeting).

Software Testing & Revision. 11 weeks has been allocated for initial testing and revision of the modified LOCKSS software in conjunction with firewall systems by the Preservation Committee. During this period, test installations of the software (both replication and harvesting capabilities) will be set up at two or more of the development sites, and the operation of the software will be evaluated. At the end of this period the software should be ready for deployment by all sites, following the network deployment plan developed at the second all-project meeting.

Initial Content Harvest. 13 weeks has been allocated for the initial content harvest. It is unlikely that the entire period will be needed for the harvest, but time has been allotted for contingencies and repetitions to correct problems that may initially occur. By the end of this time a validated content harvest of the highest priority content should have been achieved.

Subsequent Content Harvests. The next year of the project has been slated for subsequent content harvests, guided by further enhancements to the conspectus, and the subsequent harvest plan developed at the third all-project meeting.

A.6.6 Cooperative Agreement Work Plan

Cooperative Agreement Analysis. A study of cooperative agreement elements will be undertaken by the Steering Committee over a period of 39 weeks, guided by the cooperative agreement analysis plan developed at the third all-project meeting. A liberal amount of time has been allocated for this activity to enable extended discussions and observation of the dynamics of the preservation network.

Cooperative Agreement Draft. The draft cooperative agreement will be developed over a period of 26 weeks by the Steering Committee. Versions will be circulated via email and discussed in periodic conference calls.

Steering Committee Meets. This meeting will focus on drafting the cooperative agreement. *Meeting Activities:* Review the results of A) the cooperative agreement analysis, B) the enhanced conspectus, C) results from subsequent content harvests, and D) early results from tests of the preservation network. *Meeting Outcomes:* 1) Consensus on key elements of the cooperative agreement, 2) outline of the cooperative agreement, and 3) a timetable for next steps to develop the cooperative agreement.

Cooperative Agreement Completed. An additional 25 weeks has been allocated for completion of all aspects of the cooperative agreement. This includes finalizing the document(s) comprising the agreement, formation of any organizational entities that may be required, and any other work that may be required.

Steering Committee Meets. This meeting will be aimed at finalizing the cooperative agreement. *Meeting Activities:* Review and discuss: A) the final cooperative agreement to resolve any remaining issues of significance, B) the results of the preservation network tests, and C) the final project report. *Meeting Outcomes:* 1) Finalized cooperative agreement, 2) outline of the final project report.

A.6.7 Preservation Network Work Plan

Network Deployment. All six preservation nodes will be deployed over a period of 13 weeks. Preparations for this activity will have included ordering and receiving the equipment, software development and testing, and hiring technical staff needed for deployment at each site. All preservation nodes will be configured and thereby linked, but not yet synchronized in terms of content holdings.

Network Synchronization. Once the preservation network is deployed and configured properly, the content harvest will be disseminated throughout the network nodes. This synchronization of archived content will hopefully occur

relatively rapidly (we anticipate no more than 10 ten days, based on preliminary estimates), especially if the Internet 2 linkages between the sites can be utilized (this would enormously speed up the data transfer rate). However, a full 13 weeks has been allocated in case this synchronization takes longer than expected.

Network Testing. More than a year has been allocated for testing of the preservation network, guided by the preservation network assessment plan developed at the third all-project meeting. Various elements of reliability, security, fault tolerance and recovery processes will be tested. At least one simulated catastrophic data loss and recovery case study will be conducted.

Steering Committee Meets. The last project meeting will focus on completing the final project report. *Meeting Activities:* Review and discuss: A) the results of the continuing preservation network tests, and B) the final project report. *Meeting Outcomes:* 1) Semifinal project report, which will be finished in consultation via email and phone calls during the remaining months of the project.

A.9 Approved Schedule

The deliverables described in Exhibit A.6, the Project reports described in Exhibit A.7 and the meetings described in Exhibit A.8 are incorporated into the Schedule set forth below. Awardee will comply with the Schedule. Awardee will also prepare reports and deliver Work Plan deliverables to the Library in a timely fashion. Awardee will attend and participate in meetings and conference calls as reasonably requested by the Library.

1	Project Commencement		Wed 9/1/04	Wed 9/1/04
2	A. Content Conspectus	438d	Wed 9/1/04	Fri 5/5/06
3	A1. Content Committee Preparation	23d	Wed 9/1/04	Fri 10/1/04
4	All Project Meeting #1	1d	Fri 10/1/04	Fri 10/1/04
5	A2. Conspectus Development	100d	Mon 10/4/04	Fri 2/18/05
6	All Project Meeting #2	1d	Fri 2/18/05	Fri 2/18/05
7	A4. Conspectus Prioritization	55d	Mon 2/21/05	Fri 5/6/05
8	A5. Conspectus Enhancement	260d	Mon 5/9/05	Fri 5/5/06
9	All Project Meeting #3	1d	Tue 8/16/05	Tue 8/16/05
10	B. Content Harvest	510d	Wed 9/1/04	Tue 8/15/06
11	B1. Software Modification	130d	Wed 9/1/04	Tue 3/1/05
12	B2. Software Testing & Revision	55d	Wed 3/2/05	Tue 5/17/05
13	B3. Initial Content Harvest	65d	Wed 5/18/05	Tue 8/16/05
14	B4. Subsequent Content Harvests	260d	Wed 8/17/05	Tue 8/15/06
15	C. Cooperative Agreement	450d	Mon 5/9/05	Fri 1/26/07
16	C1. Cooperative Agreement Analysis	195d	Mon 5/9/05	Fri 2/3/06
17	C2. Cooperative Agreement Draft	130d	Mon 2/6/06	Fri 8/4/06
18	Steering Committee Meets	1d	Fri 2/3/06	Fri 2/3/06
19	C3. Cooperative Agreement Completion	125d	Mon 8/7/06	Fri 1/26/07
20	Steering Committee Meets	1d	Fri 8/4/06	Fri 8/4/06
21	D. Preservation Network	462d	Mon 5/9/05	Tue 2/13/07
22	D1. Network Deployment	65d	Mon 5/9/05	Fri 8/5/05
23	D2. Network Synchronization	65d	Wed 8/17/05	Tue 11/15/05
24	D3. Network Testing	325d	Wed 11/16/05	Tue 2/13/07
25	Steering Committee Meets	1d	Tue 2/13/07	Tue 2/13/07
26	Final Reporting and Wrap-up completed		Fri 8/31/07	Fri 8/31/07