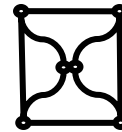

HARVEST PLAN

MetaArchive of Southern Digital Culture Project



2005-03-11

Summary

This document describes the plan for conducting the initial and subsequent harvesting activities of the MetaArchive of Southern Digital Culture Project. Harvesting activities primarily include both 1) preparatory creation of required LOCKSS manifests and 2) the activation of the harvesting process in the individual site installations. This document makes reference to the project plan and other project documents which provide more details about other aspects of the project.

Where

Harvesting activities will be conducted during Project Phase B3 (Initial Content Harvest) at *the four Development Sites identified in the project plan*: Emory, FSU, GA Tech, and VA Tech. We will endeavor to also conduct harvesting activities at Louisville and Auburn during B3 depending on how the initial harvest goes forward but, at a minimum, initial harvesting *must* occur at the Development Sites for an adequate assessment of the process. During Project Phase B4 (Subsequent Content Harvests), harvesting *must* be conducted at *all six* primary preservation sites, and *may* be conducted at the Library of Congress and other potential sites that may join the cooperative over time.

What

In order to conduct harvesting activities, several things must be accomplished:

- 1) Each of the six partner sites must create functional manifest pages and associated plug-ins for the collections selected for the initial harvest
- 2) A minimally functional preservation network must be established comprised of two or more operational preservation nodes, with the network's title database and plug-in registry populated with data associated with the collections selected for the initial harvest
- 3) The harvesting process on these servers must be aimed at the manifest pages, initiated, and successfully completed to populate the network with the content of the first harvest

Completing these steps will complete Phase B3 (Initial Content Harvest). Adding additional preservation nodes and collections will comprise the activities in Phase B4 (Subsequent Content Harvests).

Who

The respective members of the Steering Committee will be responsible for ensuring that the required activities for the initial and subsequent content harvests take place at the six primary preservation sites, according to the timetable set forth below. Technical staff working on the project may be the ones who actually do much of the work in practice, but the Steering Committee members are the individuals responsible for ensuring that the work gets done in time to meet the project timeline.

How

As described in the project plan, the Content Harvest is to occur in four phases:

Phase B1 (Software Modification) was an initial software analysis phase in which LOCKSS was jointly explored with Thomas Robertson to identify whether or not the software would require modification (no modifications were required, other than the creation of a separate instantiation of the LOCKSS network (suggestion that we call it CLOCKSS to differentiate it in our discussions).

Phase B2 (Software Testing and Revision) is the current phase in which we will begin installing, configuring, testing, and reconfiguring the software until we have confidence that it is working properly. The sites holding the collections selected for the initial harvest will have to begin setting up manifest pages and creating the required plug-ins for the initial collections to be harvested during this period. The manifests have to be placed on the servers where the collections are. The plug-ins will have to be added to the CLOCKSS plug-in registry for the MetaArchive preservation network. Finally, a title database will have to be created for the collections selected for initial harvesting. This title database is used by the interface that each site will utilize to activate the harvesting process for the initial harvest.

Phase B3 (Initial Content Harvest) is when the first harvesting activities will begin, as described above. During this period the CLOCKSS nodes will begin to harvest, replicate, and record the replication of the collections selected from the conspectus for the MetaArchive. Each site will know what has been archived at any point through examination of the node records. Please note that sites which contribute collections to the initial harvest need not necessarily bring up functioning CLOCKSS preservation nodes (and vice versa) during B3, but all sites must bring up a functioning node some time during B4.

Phase B4 (Subsequent Content Harvests) is when additional nodes and collections will be added to the preservation network.

When

The following table of information from the project plan summarizes the timetable for content harvesting phases:

PHASE	DURATION	START DATE	END DATE
B. All Content Harvest Phases	518 days	1-Sep-2004	25-Aug-2006
B1. Software Modification	130 days	1-Sep-2004	1-Mar-2005
B2. Software Testing & Revision	55 days	2-Mar-2005	17-May-2005
B3. Initial Content Harvest	65 days	30-May-2005	26-Aug-2005
B4. Subsequent Content Harvests	260 days	29-Aug-2005	25-Aug-2006

Harvest Prioritization Decisions

A discussion concerning the preliminary conspectus database produced the table below, and resulted in some decisions regarding the initial harvest. Each of the six partner sites committed to preparing at least one collection for the initial harvest, meaning that a manifest and plug-in would be prepared before August 1 by the respective sites for the six top collections identified in the table below. These collections were selected primarily as an effective initial set to use in the set-up and testing of the preservation network. The collections are of high value individually, but are not necessarily the most critical collections for preservation in every case. Rather, each institution was given discretion to choose a single collection that they would commit to preparing for harvest in this timeframe. The figures in the following table should be considered approximate estimates for planning purposes.

Conspectus ID	Collection Title	Size in Bytes (preliminary estimate)	LOCKSSable	Subject Relevance	Risk Factor	Institution	Initial Harvest
4	Alabama Cooperative Extension Service (ACES) Photographs, 1920s-1960s	18,000,000	4	5	2	Auburn	Yes
37	Southern Changes	25,000,000	4	5	2	Emory	Yes
	FSU Institutional Repository	2,000,000,000	3	4	4	FSU	Yes
31	SMARTech	10,000,000,000	4	4	5	GA Tech	Yes
12	Bernheim Foundation interviews	24,134,411,112	4	5	3	Louisville	Yes
22	University Archives of Virginia Tech	10,000,000,000	2	5	4	VA Tech	Yes
34	Glomerata: Auburn University Yearbooks, 1897-	10,000,000,000	5	5	2	Auburn	Possible
39	The Civil War in America from the Illustrated London News	3,514,536	4	4	2	Emory	Possible
19	Sam Nunn Constituent Mail System Files	859,823,104	4	5	2	Emory	Possible
20	Special Collections and Archives Digital Image Master Files	2,147,483,647	4	4	4	Emory	Possible
18	Southern Spaces	160,000,000,000	4	5	4	Emory	Possible
7	FSU Historic Photograph Collection	14,816,378	5	5	2	FSU	Possible
6	Digitized Juvenile Literature	2,147,483,647	3	1	1	FSU	Possible
24	The Buildings of Georgia Tech from 1888-1908	36,000	4	5	2	GA Tech	Possible
25	Photographs of the Historic American Buildings Survey Georgia	326,000	4	5	2	GA Tech	Possible
15	Georgia Tech Advertisements	7,025,459	4	5	2	GA Tech	Possible
27	Deceased Faculty Biographies	7,500,000	4	4	2	GA Tech	Possible
30	Georgia Tech Publications	14,000,000	3	5	2	GA Tech	Possible
16	"Splendid Growth:" Architectural Drawings of the A. French Textile Building	20,971,520	4	5	2	GA Tech	Possible
29	Georgia Tech Photograph Collection	23,000,000	4	5	2	GA Tech	Possible
26	An Illustration and Mensuration of Solid Geometry	39,000,000	4	1	2	GA Tech	Possible
11	George Griffin Photograph Collection	39,845,888	3	5	3	GA Tech	Possible
9	A Photographic Atlas of Selected Regions of The Milky Way	1,000,000,000	4	4	3	GA Tech	Possible
32	Georgia Tech ETDs	28,000,000,000	4	4	5	GA Tech	Possible
35	Jean Thomas collection	15,999,082,014	4	5	3	Louisville	Possible
21	Kentucky Quilt Project image masters	16,535,624,090	4	5	2	Louisville	Possible
5	International Archive of Women in Architecture Biographical Database	100,000,000	4	4	4	VA Tech	Possible
14	Graphic Comm Central	1,000,000,000	3	2	4	VA Tech	Possible
28	ETDs @ VT	35,000,000,000	3	4	5	VA Tech	Possible
33	VT Digital Imaging Archives	100,000,000,000	2	4	5	VA Tech	Possible
	Total all current collections:	419,136,943,395					

The partner sites will also attempt to prepare additional collections for the initial harvest, subject only to the current individual institutional quota of approximately 400 GB. Rephrased, institutions will refrain from adding a particular additional collection to the MetaArchive if that addition would put the overall institutional contribution to the entire system above 400 GB.

The Steering Committee further agreed that the quota arrangements will be discussed in subsequent project meetings, to see if they should be renegotiated using a more complex system reflecting both collection value and other factors. The group recognized that the question of institutional quotas is complicated, and should be carefully considered in the overall context of cooperative agreement analysis.

Next Steps Targets

Plug-ins and manifests: After the collections for the initial harvest are selected from the conspectus at the March 11 all-project meeting, the Steering Committee members at the relevant institutions need to begin the process of creating the plug-ins and manifests. Everyone should aim to create these and test them by June 1. The drop-dead date for finishing these is August 1, in order to allow adequate time to complete the initial harvest. Put differently, whatever plug-ins and manifests are working by August 1 will effectively define the initial harvest, which needs to be completed by August 26.

Preservation Nodes: The Development Sites should try to get the CLOCKSS servers running by June 15. Drop-dead is August 1 to participate in the initial harvest. See the Network Deployment Plan for more details.

Notes on this document

This document was completed at the second MetaArchive All-Project Meeting held at Emory University on Friday 2005-03-11, by the Steering Committee and additional attending project participants. The document was last updated by Martin Halbert on 2005-03-14.