

Review of Multi-Institutional Repository

Erica Wilson

Digital Preservation

Howard Besser
&
Chris Lacinak

Today more than 93 percent of the world's information is produced as digital files and not print documents. From sound and moving images, emails to government documents, scanned images to those born digital, how does one preserve them? In addition to natural disasters and power failures there is also the threat of hardware and software failure or corruption. Through this need multiple organizations and groups are creating way to preserve these sources of information. One such organization is the MetaArchive Cooperative. ¹

The MetaArchive Cooperative provides an affordable preservation services to museums, universities, libraries, and other cultural heritage institutions. Six academic libraries and the library of congress founded the MetaArchive in 2004. Their plan was to develop an organization for the preservation and distribution of digital materials. MetaArchive Cooperative promotes and supports their collaborative approach to distribute digital preservation practices in 2006 it formed its own non-profit, unincorporated, international management organization, the Educopia institute. The Educopia institute provides administrative, fiscal and strategic oversight for the MetaArchive.

The cooperative functions as a community owned and led initiative. Its collaborative network consist of several multi cultural institutions and seeks to cooperative preserve their digital material. One thing that makes MetaArchive so different is that they do not out source to others but each member actively participates in preserving their own material. Currently the Cooperative has eight sustaining members and over six-dozen organizations that are working towards membership. The network is currently preserving over 100 different collections and anticipates to continue doubling in size each year. (this is numbers from 2008)².

The Cooperative provides its members with its services through decentralized, geographically dispersed organizational and technical framework. This is designed to promote group efforts in cooperating organizations that seek to actively preserve their digital collection. The MetaArchive does not have a single

¹ . <<http://www.metaarchive.org/>>

² MetaArchive Cooperative Charter. Atlanta: Educopia Institute, 2009

location of operation and does not control production in of its services in one central way. Additionally, MetaArchive owns no equipment or other assets.

MetaArchive differs significantly from other digital preservation networks in that they practice distributed digital preservation. Distributed digital preservations is when multiple copies are store in a variety of different geographical locations, using a technology infrastructure based on LOCKSS (Lots Of Copies Keeps Stuff Safe) software developed at Stanford University. The members provide secure servers called, MetaArchive-LOCKSS Caches, in their own setting. The success of the MetaArchive depends on the individual members participation.³

The MetaArchive is a true cooperative of contributing institutions and it depends on its members to participate in the preservation activities and to shoulder the preservation responsibilities. The Educopia management organization helps to level the playing field between contributors and no one organization plays a dominant role based on its size or level of resources. Every member of the cooperative pays the same membership fees and has the same number of representatives on the committee.⁴

The MetaArchive runs on what is called “dark archives” this means that no access is provided to non-members. The networks perform bit-level preservation for all file types and preservation with migration. The foundation for MetaArchive distribution technical system is the open source LOCKSS. The Cooperatives network is made up of servers that ingest content and preserve them, these networks are called LOCKSS Caches. The technical framework of MetaArchive allows the LOCKSS Caches to constantly check in with each other for signs of “bit rot” or cache/file corruption. Each Cache is under control by separate administrative organizations. This is a way to guard against natural and technical disaster. This system provides a means of constantly checking each file and provides repairs when necessary.⁵

³ MetaArchive Cooperative Charter. Atlanta: Educopia Institute, 2009

⁴ MetaArchive web page, 10 Oct. 2010. <<http://www.metaarchive.org/>>

⁵ Skinner, Katherine and Matt Schultz, eds. A guide to digital preservation. Atlanta: Educopia Institute, 2010.

Since the beginning MetaArchive Cooperative has established a variety of genre and subject based archives within the network. One example is the “Southern Digital Culture.” It was an archive set up to collect music, videos, photos, artwork, text, websites and other digital materials all related to the history of the south. In 2008 they launched their “electronic Theses and Dissertations” archive in cooperation with the NDLTD in an effort to preserve these items with its participation members. The MetaArchive also has an archive network established boundaries for non subject or genre based to encompass content from members that may not fit into an existing subject/ genre based archive. This is so that if in the future they decide to create a specific archive for a subject, which in the past was not in demand, they have the information saved and ready to go. For example they are working on an archive for digital newspapers to meet the demands of the new and existing members. A large amount of information is already collected on digital newspapers and stored.⁶

When Cultural memory organizations collaborate they tend to be grant funded and only for two to three years. Collaborating on a digital front requires a much different kind of relationship, one with long-term goals and a formal infrastructure. When institutions join together in managing Private LOCKSS Networks (PLN) they need to first decide and think about what kind of organizational framework they need to function in. The distributed digital preservation solution (DDP) requires that the collaborating institutions agree to preserve each other’s content. Along with this, issues around roles and responsibilities of contributors and government structures all need to be considered. Each governing structure needs to address questions in regards to management of members and services. These questions include: who can join the organization, what are their responsibilities, what model is used to lead the group, how is the governing body organized, can memberships be revoked, etc.⁷

⁶ MetaArchive web page, 10 Oct. 2010. <<http://www.metaarchive.org/>>

⁷ Skinner, Katherine and Matt Schultz, eds. A guide to digital preservation. Atlanta: Educopia Institute, 2010.

The field of distributed digital preservation is still very new. There are very few successful DDP models available. MetaArchive has a content committee that is comprised of members of the organization to propose new subject and general based archives to the Steering committee. The members of the MetaArchive decide on what projects to take on.

The Cooperative differs in business in that they provide a way for the cultural memory institutions to own and control the process of digital preservation for themselves. MetaArchive encourages organizations to invest in their own infrastructures and build their own capabilities by participating in the decentralized preservation network.⁸

Determining the groups decide upon the scope of materials covered shared collection and the individual institutions. For example the group took a survey or poll and decided to work on the theses and dissertation preservation project as well as the digital newspapers. Yet when an individual organization wanted to do an archive on the American south the group agreed but left the scope very broad to include less subject driven materials to be gathered for other digital archives. Some materials that may not strictly fit the criteria of the network but warrant preservation are still collected by keeping the scope broad although, sometimes they may choose to narrow the scope by restricting it to a specific time period or geographical region. Once an archive has set its scope it is important to document what content is on target and to discuss the possibility of later broadening the scope. This is to help the network grow as well as attract new members.⁹

LOCKSS performs a web-based ingestion of targeted content that is made available by the contributing institutions. Content that is slated for ingest must, be available at a web accessible location, organized and documented with correct metadata and promote human understanding of the content that is being preserved, plugin creation, manifest page and able to differentiate structurally from other

⁸ MetaArchive Cooperative Charter. Atlanta: Educopia Institute, 2009

⁹ Skinner, Katherine and Matt Schultz, eds. A guide to digital preservation Atlanta: Educopia Institute, 2010.

content hosted by the content contributor. In order to ingest digital content the LOCKSS cache need to know it has permission to ingest the information by providing the LOCKSS software information about the collection with clearly defined parameters for what the caches will use when ingesting the information. This is done with a manifest page with links for what is to be ingested or a plugin for each collection is provided. Plugins provide instructions for the software on what to ingest. When a content contributor submits a collection for ingestion the system administrator for each cache must be notified of it existence to trigger the ingestion process. When a LOCKSS daemon uses a plugin to conduct an initial harvest, it crawls the content site exhaustively within the boundaries set by the plugin parameters. Once all the content is ingested some one uses the monitoring tools on each cache to ensure the ingestion was completed successfully on at least seven caches and that all caches contain identical copies of each file.

The Conspectus Tool is a web based data management tool that maintains both LOCKSS specific metadata and descriptive collection level metadata about each digital collection that is submitted for ingestion. The conspectus contains metadata schema designed by the cooperative that includes elements from widely used metadata standards like Dublin Core, METS and MODs.

Each LOCKSS cache revisits the site from which it originally ingested a collection to check for changes and additions to content. It will then re ingest digital object from the site, not replacing the original but adding a new dated entry that contains any changes that may have occurred. This is helpful in that most collections do change over time. This helps to insure that valuable content is not lost over time.¹⁰

A key difference between traditional and digital preservation is that digital preservation needs to happen early in the digital objects life to make it viable. A book can be repaired and salvaged even after the damage has begun but just the slightest corruption in a digital file may make it completely unsalvageable. Once

¹⁰ Skinner, Katherine and Matt Schultz, eds. A guide to digital preservation Atlanta: Educopia Institute, 2010.

digital information is created it cannot be ignored for decades. File formats hardware and software degrades and become obsolete rendering this information inaccessible.

LOCKSS software provides bit-level preservation for digital objects of any file format. Bit level preservation should be present in any digital preservation solution. Basically it ensures that all the 1's and 0's that make up a file are kept intact. LOCKSS provides a set of services that make a preserved file accessible and usable in the future. In 2005 LOCKSS created and tested format migration strategies that are applied in the network. Several PLNs that are now beginning to create migration pathways for a file type and implement it in their LOCKSS environment. This will help to ensure that the MetaArchive Cooperative will continue to have access their information collected in their archives in the future.¹¹

¹¹ Skinner, Katherine and Matt Schultz, eds. A guide to digital preservation Atlanta: Educopia Institute, 2010.

Bibliography/webography

McMillan, Gail and Katherine Skinner. NDLTD/ MetaArchive

Preservation Strategy. 3rd ed. Atlanta: Educopia Institute, June
2010.

MetaArchive Cooperative Charter. Atlanta: Educopia Institute, 2009.

MetaArchive web page, 10 Oct. 2010. <<http://www.metaarchive.org/>>

Skinner, Katherine and Matt Schultz, eds. A guide to digital preservation.

Atlanta: Educopia Institute, 2010.