Benign Neglect: Developing Life Rafts for Digital Content

In his keynote speech at the Archiving 2009 Conference in Arlington, Virginia, Clifford Lynch called for the development of a benign neglect model for digital preservation, one in which as much content as possible is stored in whatever manner available in hopes of there someday being enough resources to more properly preserve it. This is an acknowledgment of current resource limitations relative to the burgeoning quantities of digital content that need to be preserved. We need low cost, scalable methods to store and preserve materials. Over the past few years, a tremendous amount of time and energy has, sensibly, been devoted to developing standards and methods for best practices. However, a short survey of some of the leading efforts clarifies for even the casual observer that implementation of the proposed standards is beyond many of those who are creating or hosting digital content, particularly because of restrictions on acceptable formats, requirements for extensive metadata in specific XML encodings, need for programmers for implementation, costs for participation, or simply a lack of a clear set of steps for the uninitiated to follow (examples include: Planets, PREMIS, DCC, CASPAR, iRods, Sound Directions, HathiTrust).[1] The deluge of digital content coupled with the lack of funding for digital preservation and exacerbated by the expanding variety of formats, makes the application of extensive standards and extraordinary techniques beyond the reach of the majority. Given the current circumstances, Lynch says, either we can seek perfection and store very little, or we can be sloppy and preserve more, discarding what is simply intractable.[2]

In contrast, other leaders of the digital preservation movement have been stating for years that benign neglect is not a workable solution for digital materials. Eric Van de Velde, director of Caltech's Library Information Technology Group, stated that the "digital archive must be actively managed."[3] Tom Cramer of Stanford University agrees: "Benign neglect doesn't work for digital objects. Preservation requires active, managed care."[4] The Digital Preservation Europe website argues that benign neglect of digital content "is almost a guarantee that it will be inaccessible in the future."[5] Abby Smith goes so far as to say that "neglect of digital data is a death sentence."[6]

Arguments to support this statement are primarily those of media or data carrier storage fragility and obsolescence of hardware, software, and format. However, the impact of these arguments can be reduced to a manageable nightmare. By removing as much as possible of the intermediate systems, storing open-source code for the software and operating system needed for access to the digitized content, and locating archival content directly on the file system itself, we reduce the problems to primarily that of format obsolescence. This approach will enable us to forge ahead in the face of our lack of resources and our rather desperate need for rapid, cheap, and pragmatic solutions.

Current long-term preservation archives operating within the Open Archival Information System (OAIS) model assume that producers can meet the requirements of ingest.[7] However, the amount of content that needs to be deposited into archives and the expanding variety of formats and genres that are unsupported, are overwhelming the ability of depositors to prepare content for preservation. Andrea Goethals of Harvard proposed that we revisit assumptions of producer ability to prepare content for deposit in accordance with the current best practices.[8]

For those producers of content who are not able to meet the requirements of ingest, or who do not have access to an OAIS archive provider, what are the options? With the recent downturn in the economy, the availability of staff and the funding for the support of digital libraries has no doubt left many collections at risk of abandonment. Is there a method for preparation of content for long-term storage that is within the reach of existing staff with few technical skills? If the content cannot get to the safe harbor of a trusted digital library, is it consigned to extinction? Or are there steps we can take to mitigate the potential loss?

The OAIS model incorporates six functional entities: ingest, data management, administration, preservation planning, archival storage, and access.[9] Of these six, only archival storage is primary; all the others are useless without the actual content. And if the content cannot be accessed in some form, the storage of it may also be useless. Therefore the minimal components that must be met are those of archival storage and some form of access. The lowest cost and simplest option for archival storage currently available is the distribution of multiple copies dispersed across a geographical area, preferably on different platforms, as recommended by the current LOCKSS initiative,[10] which focuses on bit-level preservation.[11] Private LOCKSS Network models (such as the Alabama Digital Preservation Network)[12] are the lowest-cost implementation, requiring only hardware, membership in LOCKSS, and a small amount of time and technical expertise.

Reduction of the six functional entities to only two negates the need

Jody L. DeRidder (jlderidder@ua.edu) is Head, Digital Services, University of Alabama.

for a tremendous amount of metadata collection. Where the focus has been on what is the best metadata to collect, the question becomes: What is the minimal metadata and contextual information needed? The following is an attempt to begin this conversation in the hope that debate will clarify and distill the absolutely necessary and specific requirements to enable long-term access with the lowest possible barrier to implementation. If we consider the purpose of preservation to be solely that of ensuring long-term access, it is possible to selectively identify information for inclusion. The recent proposal by the researchers of The National Geospatial Digital Archive (NGDA) may help to direct our focus. They have defined three architectural design principles that are necessary to preserve content over time: the fallback principle, the relay principle, and the resurrection principle.[13]

In the event that the system itself is no longer functional, then a preservation system should support some form of hand-off of its content—the fallback principle. This can be met by involvement in LOCKSS, as specified above. Lacking the ability to support even this, current creators and hosts of digital content may be at the mercy of political or private support for ingest into trusted digital repositories.[14] The recently developed BagIt File Package Format includes valuable information to ensure uncorrupted transfer for incorporation into such an archive.[15] Each base directory containing digital files is considered a bag, and the contents can be any types of files in any organization or naming convention; the software tags the content (or payload) with checksums and manifest, and bundles it into a single archive file for transfer and storage. An easily usable tool to create these manifests has already been developed to assist underfunded cultural heritage organizations in preparing content for a hosting institution or government infrastructure willing to preserve the content.[16] The gap of who would take and manage the content is still uncertain.

The relay principle states that a preservation system should support its own migration. Preserving any type of digital information requires preserving the information's context so that it can be interpreted correctly. This seems to indicate that both the intellectual context and the logical context need to be provided. Context may include provenance information to verify authenticity, integrity, and interpretation;[17] it may include structural information about the organization of the digital files and how they relate to one another; and it should certainly include documentation about why this content is important, for whom, and how it may be used (including access restrictions).

Because the cost of continued migration of content is very high, a method of mitigating that cost is to allow content to become obsolete but to support sufficient metadata and contextual information to be able to resurrect full access and use at some future time—the resurrection principle. To be able to resurrect obsolete materials, it would be advisable to store the content with open-source software that can render it, an open-source operating system that can support the software, and separate plain-text instructions for how to reconstruct delivery. In addition, underlying assumptions of the storage device itself need to be made explicit if possible (type of file system partition, supported length of file names, character encodings, inode information locations, etc.). Some of the need for this form of preservation may be diminished through such efforts as the Planets TimeCapsule Deposit.[18] This consortium has gathered the supporting software and information necessary to access current common types of digital files (such as PDF), for long-term storage in Swiss Fort Knox.

One of the drawbacks to gathering and storing content developed during digitization is that developing digital libraries usually have a highly chaotic disorganization of files, directory structures, and metadata that impede digital preservation readiness.[19] If the archival digital files cannot be easily and readily associated with the metadata that provides their context, and if the files themselves are not organized in a fashion that makes their relationships transparent, reconstruction of delivery at some future point is seriously in question. Underfunded cultural heritage institutions need clear specifications for file organization and preparation that they are capable of meeting without programming staff or extensive time commitments. Particularly in the current economic downturn, few institutions have the technical skills to create METS wrappers to clarify file relationships.[20]

One potential solution is to use the organization of files in the file system itself to communicate clearly to future archivists how the files relate to one another. At the University of Alabama, we have adopted a standardized file naming system that organizes content by the holding institution and type, collection, item, and then sequence of delivery (see figure 1). The file names are echoed in the file system: top level directories match the holding institution number sequence, secondary level directory names match the assigned collection number sequence, and so forth.

Metadata and documentation are stored at whatever level in the file system corresponds to the files to which they apply, and these text and XML files have file names that also correspond to the files to which they apply, which assists further in identification (see figure 2).[21]

By both naming and ordering the files according to the same system, and bypassing the need for databases, complex metadata schemes and software, we leverage the simplicity of the file system to bring order to chaos and to enable our content to be easily reconstructed by future systems.

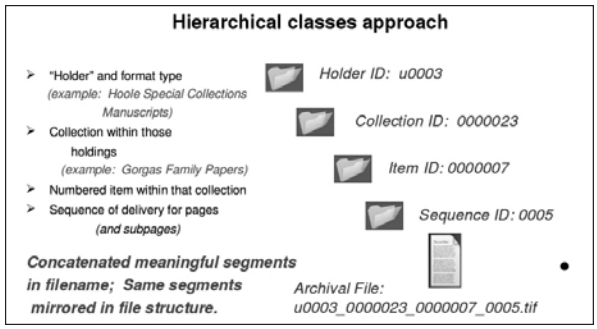

Figure 1. University of Alabama Libraries Digital File Naming Scheme (©2009. Used with permission.)

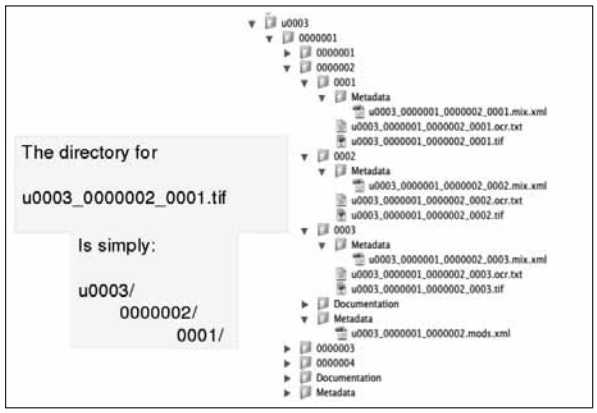

Figure 2. University of Alabama Libraries Metadata Organization (©2009. Used with permission.)

While no programmers are needed to organize content into such a clear, consistent, and standardized order, we are developing scripts that will assist others who seek to follow this path. These scripts not only order the content, they also create LOCKSS manifests at each level of the content, down to the collection level, so that the archived material is ready for LOCKSS pickup. A standardized LOCKSS plugin for this method is available.

To assist in providing access without a storage database, we are also developing an open-source web delivery system (Acumen),[22] which dynamically collects content from this protected archival storage arrangement (or from web-accessible directories) and provides online delivery of cached derivatives and metadata, as well as webcrawler-enabled content to expand accessibility. This model of online delivery will enable low cost, scalable development of digital libraries by simply ordering content within the archival storage location.

Providing simple, clear, accessible methods of preparing content for preservation, of duplicating archival treasures in LOCKSS, and of web accessibility without excessive cost or deep web database storage of content, will enable underfunded cultural heritage institutions to help ensure that their content will continue to survive the current preservation challenges. As David Seaman pointed out, the more a digital item is used, the more it is copied and handled, the more it will be preserved.[23] Focusing on archival storage (via LOCKSS) and accessibility of content fulfills the two most primary OAIS functional capabilities and provides a life raft option for those who are not currently able to surmount the forbidding tsunami of requirements being drafted as best practices for preservation.

The importance of offering feasible options for the continued support of the long tail of digitized content cannot be overstated. While the heavily funded centers may be able to preserve much of the content under their purview, this is only a small fraction of the valuable digitized material currently facing dissolution in the black hole of our cultural memory. As Clifford Lynch pointed out, funding cutbacks at the sub-federal level are destroying access and preservation of government records; corporate records are winding up in the trash; news is lost daily; and personal and cultural heritage materials are disappearing as we speak.[24] It is valuable and necessary to determine best practices and to seek to employ them to retain as much of the cultural and historical record as possible, and in an ideal world, these practices would be applied to all valuable digital content. But in the practical and largely resource-constrained world of most libraries and other cultural institutions, this is not feasible. The scale of content creation, the variety and geographic dispersal of materials, and the cost of preparation and support makes it impossible for this level of attention to be applied to the bulk of what must be saved. For our cultural memory from this period to survive, we need to communicate simple, clear, scalable, inexpensive options to digital holders and creators.

References

1. Planets Consortium, Planets Preservation and Long-Term Access Through Networked Services, http://www.planets-project.eu/ (accessed Mar. 29, 2011); Library of Congress, PREMIS (Preservation Metadata Maintenance Activity), http://www.loc.gov/standards/premis/ (accessed Mar. 29, 2011); DCC (Digital Curation Centre), http://www.dcc.ac.uk/ (accessed Mar. 29, 2011); CASPAR (Cultural, Artistic, and Scientific Knowledge for Preservation, Access, and Retrieval), http://www.casparpreserves.eu/ (accessed Mar. 29, 2011); IRODS (Integrated Rule-Oriented Data System), https://www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems (accessed Mar. 29, 2011); Mike Casey and Bruce Gordon, Sound Directions: Best Practices for Audio Preservation, http://www.dlib.indiana.edu/projects/sounddirections/papers Present/sd_bp_07.pdf (accessed June 14, 2010); HathiTrust: A Shared Digital

Repository, http://www.hathitrust.org/ (accessed Mar. 29, 2011).

2. Clifford Lynch, *Challenges and Opportunities for Digital Stewardship in the Era of Hope and Crisis* (keynote speech, IS&T Archiving 2009 Conference, Arlington, Va., May 2009).

3. Jane Deitrich, e-Journals: Do-It-Yourself Publishing, http://eands.caltech.edu/articles/E%20journals/ejournals5.html (accessed Aug. 9, 2009).

4. Tom Cramer, quoted in Art Pasquinelli, "Digital Libraries and Repositories: Issues and Trends" (Sun Microsystems presentation at the Summit Bibliotheken, Universitäsbibliothek Kassel, 18–19, Mar. 2009), slide 12, http://de.sun.com/sunnews/events/2009/bibsummit/pdf/2-art-pasquinelli.pdf (accessed July 12, 2009).

5. Digital Preservation Europe, What is Digital Preservation? http://www.digitalpreservationeurope.eu/what-is-digital-preservation/ (accessed June 14, 2010).

6. Abby Smith, "Preservation," in Susan Schreibman, Ray Siemens, John Unsworth, eds., *A Companion to Digital Humanities* (Oxford: Blackwell, 2004), http://www.digitalhumanities.org/companion/ (accessed June 14, 2010).

7. Consultative Committee for Space Data Systems, Reference Model for an Open Archival System (OAIS), *CCSDS 650.0-B-1 Blue Book*, Jan. 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed June 14, 2010).

8. Andrea Goethals, "Meeting the Preservation Demand Responsibly = Lowering the Ingest Bar?" *Archiving 2009* (May 2009): 6.

9. Consultative Committee for Space Data Systems, Reference Model.

10. Stanford University et al., Lots Of Copies Keep Stuff Safe (LOCKSS), http://www.lockss.org/lockss/Home (accessed Mar. 29, 2011).

11. David S. Rosenthal et al., "Requirements for Digital Preservation Systems: A Bottom-Up Approach," *D-Lib Magazine* 11 (Nov. 2005): 11, http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html (accessed June 14, 2010).

12. Alabama Digital Preservation Network (ADPNet), http://www.adpn.org/ (accessed Mar. 29, 2011).

13. Greg Janée, "Preserving Geospatial Data: The National Geospatial Digital Archive's Approach," *Archiving 2009* (May 2009): 6.

14. Research Libraries Group/OCLC, Trusted Digital Repositories: Attributes and Responsibilities, http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf (accessed July 17, 2009).

15. Andy Boyko et al., The BagIt File Packaging Format (0.96) (NDIIPP Content Transfer Project), http://www.digitalpreservation.gov/library/resources/tools/docs/bagitspec.pdf (accessed July 18, 2009).

16. Library of Congress, BagIt Library, http://www.digitalpreservation.gov/partners/resources/tools/index.html#b (accessed June 14, 2010).

17. Andy Powell, Pete Johnston, and Thomas Baker, "Domains and Ranges for DCMI Properties: Definition of the DCMI term Provenance," http://dublincore.org/documents/domain-range/index.shtml#ProvenanceStatement (accessed July 18, 2009).

18. Planets Consortium, Planets Time Capsule—A Showcase for Digital Preservation, http://www.ifs.tuwien.ac.at/dp/timecapsule/ (accessed June 14, 2010).

19. Martin Halbert, Katherine Skinner, and Gail McMillan, "Avoiding the Calf-Path: Digital Reservation Readiness for Growing Collections and Distributed Preservation Networks," *Archiving 2009* (May 2009): 6.

20. Library of Congress, Metadata Encoding and Transmission Standard (METS), http://www.loc.gov/standards/mets.

21. Jody L. DeRidder, "From Confusion and Chaos to Clarity and Hope," in *Digitization in the Real World: Lessons Learned from Small to Medium-Sized Digitization Projects*, ed. Kwong Bor Ng and Jason Kucsma, (Metropolitan New York Library Council, N.Y., 2010).

22. Tonio Loewald and Jody DeRidder, "Metadata In, Library Out. A Simple, Robust Digital Library System," *Code4Lib Journal* 10 (2010), http://journal.code4lib.org/articles/3107 (accessed Aug. 29, 2010).

23. David Seaman "The DLF Today" (keynote presentation, 2004 Symposium on Open Access and Digital Preservation, Atlanta, Ga.), paraphrased by Eric Lease Morgan in Musings on Information and Librarianship, http://infomotions.com/musings/open-access-symposium/ (accessed Aug. 9, 2009).

24. Lynch, *Challenges and Opportunities*.