CASE 11

Preserving the H-Net Academic Electronic Mail Lists

AUTHOR: **LISA M. SCHMIDT**
Electronic Records Archivist
Michigan State University
lisa.schmidt@matrix.msu.edu

PAPER DATE: February 2009

CASE STUDY DATE: 2007–ongoing

ISSUE: Assessing the existing state of preservation for the H-Net e-mail lists using digital preservation theory and the Trusted Repositories Audit & Certification: Criteria and Checklist (TRAC) evaluation tool. Making recommendations and overseeing the implementation of improvements to make H-Net a trusted digital repository. Ensuring authenticity is the primary preservation issue.

KEYWORDS: Electronic mail preservation, Electronic mail list preservation, Trusted digital repository, Data authenticity issues, Digital preservation

Background

H-Net: Humanities and Social Sciences Online is an international consortium of scholars and teachers with a mission to "create electronic networks and resources dedicated to advancing research, teaching, learning, public outreach, and professional service within their own specialized areas of knowledge."[1] Since its 1992 beginnings as a virtual service hosted at the University of Illinois at Chicago, H-Net has grown to include more than 180 scholarly social sciences and humanities networks. It currently is hosted by MATRIX: Center for Humane Arts, Letters and Social Sciences Online at Michigan State University.

While H-Net includes humanities-related reviews as well as job and meeting announcements, the heart of the consortium is its 185 free interactive discussion networks, or electronic mail lists. More than 450 editors and nearly 125,000 members participate in these networks, which surpassed one million messages as of January 2008. In the month of November 2008 alone, editors posted more than 5400 subscriber messages to the public list; an estimated 84,000 messages were viewed during the last week of that same month.

A 12-14 member council governs the policies and activities of H-Net. In addition to its public lists, H-Net includes more than 230 "private" lists used by editors, council members, and administrators for planning, testing, and advisory purposes. All of the H-Net e-mail lists run on the proprietary L-Soft LISTSERV software.

In 2007, MATRIX received a two-year grant from the National Historical Publications and Records Commission (NHPRC) to advance the state of e-mail preservation by assessing and improving upon the digital preservation practices for the H-Net electronic mailing lists in order to ensure longevity of the content. H-Net represents a compilation of years of academic discourse, with messages bookmarked and cited in scholarly research and publications. Long-term preservation of this valuable scholarly resource is important to students and practitioners in the represented academic areas, offering the potential to provide a deeper understanding of the context and evolution of their fields.

The Center for Research Libraries (CRL) and Online Computer Library Center (OCLC) Trusted Repositories Audit & Certification: Criteria and Checklist (TRAC)[2] was the primary assessment tool used to evaluate H-Net as a preservation system. Lisa Schmidt, electronic records archivist and project manager at MATRIX, directs the project, with systems administrator Dennis Boone assisting with technical matters. Other MATRIX and H-Net administrators were consulted as needed during the TRAC evaluation. Before

[1] H-Net: Humanities and Social Sciences Online, "H-Net Mission Statement," March 2000, http://www.h-net.org/about/mission.php. Retrieved 21 October 2008.

[2] The Center for Research Libraries (CRL) and Online Computer Library Center Inc. (OCLC), "Trustworthy Repositories Audit & Certification: Criteria and Checklist," Version 1.0, February 2007, http://www.crl.edu/PDF/trac.pdf.

undertaking the formal assessment, MATRIX identified the failure to ensure authenticity of messages as a major issue in the preservation of the H-Net lists.

Case Methodology

The assessment began with an examination of H-Net message posting, storage, and retrieval processes. As H-Net is hosted by MATRIX, a review of the digital humanities research center's backup and storage practices was also conducted. H-Net was then evaluated as a preservation system by applying the Open Archival Information System (OAIS) model, a reference model for an archive that has accepted the responsibility to preserve information for a designated community;[3] the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) authenticity measures;[4] and as a trusted digital repository, with TRAC used as the primary assessment tool.

How H-Net Works

The LISTSERV software allows H-Net subscribers and editors to create messages using a web browser interface or LISTSERV commands. All messages sent to a public list must be written in plain text formats, such as ASCII and Unicode, and attachments are not allowed. Copyright for a message is retained by the author; however, sending a message to a list grants H-Net and subscribers the permission to electronically distribute and download it for nonprofit educational purposes, provided that proper attribution is given to the original author, list, and date of posting.[5] When a subscriber sends a message to a list, it is first delivered to that list's editor for approval. The editor either sends the message on for posting as received or edits the content before posting the message. (See Figure 1.)

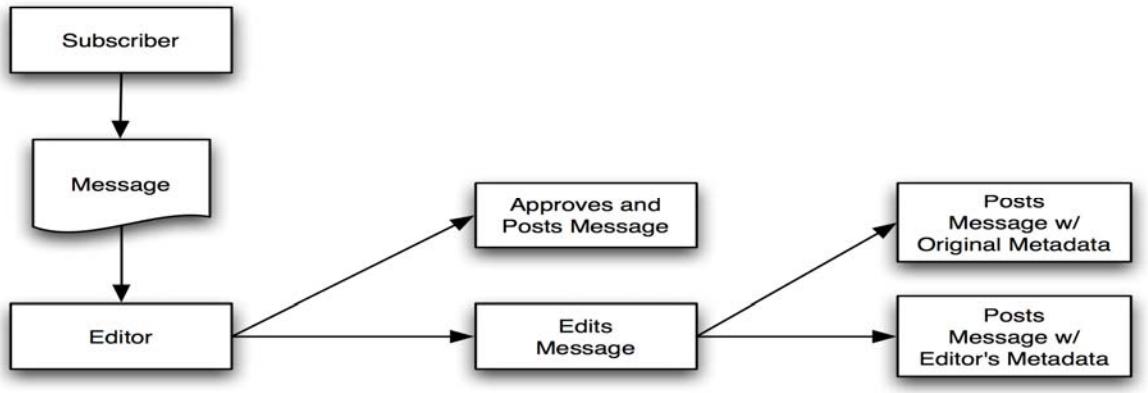

Figure 1. H-Net Message Posting Process

[3] Consultative Committee for Space Data Systems (CCSDS), "Reference Model for an Open Archival Information System (OAIS)," Blue Book 1, Issue 1, CCSDS Secretariat, January 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf.

[4] International Research on Permanent Authentic Records in Electronic Systems (InterPARES), http://www.interpares.org.

[5] H-Net: Humanities and Social Sciences Online, About H-Net, "H-Net's Policy on Copyright and Intellectual Property," 8 November 1999, http://www.h-net.org/about/intellectualproperty.php. Retrieved 24 October 2008.

A message sent for posting is added to an open "notebook" file. Each notebook contains a concatenation of messages posted in a seven-day period, in the original order that they were received. The name of a notebook file includes the name of the H-Net list, the year and month that the message posted, and a letter referencing the time period of the posting within the month ("a" for days 1-7, "b" for days 8-14, and so on). For example, a notebook with the filename "h-africa.log0802a" would include messages posted to the H-Africa list during the first seven days of February 2008.

Every 24 hours, the newest messages in the notebook files are parsed and copied to a Bibliographic Retrieval Services (BRS) database system. As a separate operation, a log browse application reads the notebook file, extracts key metadata, and creates MD5 hash algorithms for each message. A cache builder script then writes the message metadata to a MySQL database cache.

From the H-Net website, a researcher may navigate to the H-Net discussion list of choice, from which he or she may view the discussion logs for that list by month. The researcher then selects a message to view. Using keywords, the researcher also has the option to conduct a full-text search and select a message from the view provided; searches go through the BRS database. In either case, the log browse application builds a URL for the selected message and pulls it from the pertinent notebook file for display. This URL will incorporate a combination of the message's filename and MD5 hash, resulting in a unique, persistent identifier for the message that may be bookmarked and used in citations for published works. (See Figure 2.)

Systems Configuration, Backup, and Storage

MATRIX runs its operations, including the H-Net e-mail lists, on several servers kept in a climate controlled, physically secured room; these servers run the Debian distribution of Linux. As of February 2009, approximately 2.7 TB of data was stored on the servers. Incremental tape backups are performed daily, with a full backup performed on a weekly basis. Those tapes are taken to the MSU Computer Center and swapped out for the tapes that had been stored there the previous week. Backup tapes cycle through the system approximately every six weeks and are replaced as needed, such as when a cartridge breaks.

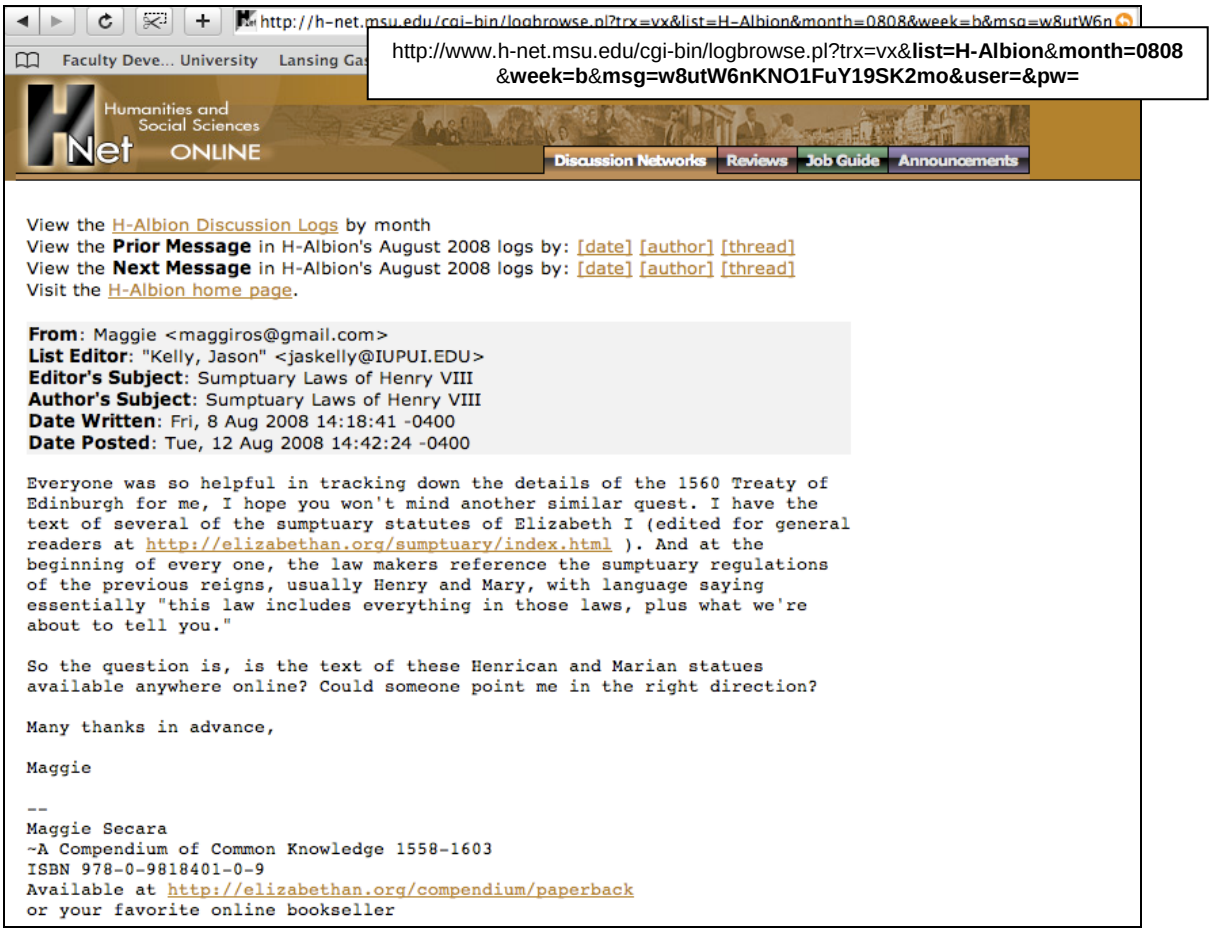

Figure 2. H-Net Message Retrieval View, with URL Detail

In addition to these ongoing backups, a full "permanent" backup is performed every one-to-two months in order to ensure against data loss. Those tapes are kept in a cabinet in a minimally secured room, presumably in perpetuity. The MATRIX systems administrator keeps a wiki-based log of all tape backups.

TRAC Audit

H-Net was evaluated as a preservation system using the Trusted Repositories Audit & Certification: Criteria and Checklist (TRAC). With a foundation in the framework presented in the Research Libraries Group (RLG)/OCLC joint publication *Trusted Digital Repositories: Attributes and Responsibilities*,[6] this checklist may be used as a "tool for objective evaluation" of a repository, whether performed in-house as a self-assessment (as has been the case with the H-Net preservation system) or by a third-party auditor. A "gap analysis" between current strategies and the desired TRAC objectives enables the

[6] Research Libraries Group (RLG), *Trusted Digital Repositories: Attributes and Responsibilities*, RLG/OCLC, May 2002, http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf.

manager of the archive to formulate the strategies necessary to close the gaps and improve the trustworthiness of the digital repository.[7]

The checklist is organized into three sections: Organizational Infrastructure; Digital Object Management; and Technologies, Technical Infrastructure, and Security. Each section includes a number of core audit criteria of a trusted digital repository for comparison to local capabilities. Examples of documentation and other evidence that show how the repository meets the criteria are also provided.[8] Each criterion on the checklist leaves space for the auditor to note how the repository meets that requirement and provides any other relevant information. (See Figure 3.)

Trustworthy Repositories Audit & Certification: Criteria Checklist

Organization:		Auditor:		Page	
Section:	A. Organizational Infrastructure	Interviewee(s):		Date	
Aspect:	A1. Governance & organizational viability				

Criterion	Evidence (Documents) Examined	Findings and Observations	Result
A1.1. Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information.			
A1.2. Repository has an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.			

Figure 3. Sample Page of TRAC Checklist [9]

While the criteria appropriate to a given repository will depend on features unique to that archive, the TRAC offers a list of minimum required documents that will satisfy multiple requirements.[10] TRAC provides a practical means to assess a repository's compliance with key features of the OAIS model and other archival preservation guidelines, as well as the administrative functions necessary for a digital repository to be deemed trusted.

At the time of the assessment, H-Net proved adequate in many areas, including governance and organizational viability, structure, and staffing; procedural accountability and policy framework; financial sustainability; contracts, licenses, and liabilities;

[7] Regents of the University of Michigan, Digital Preservation Management Workshop, November 2008.
[8] CRL and OCLC, p. 5.
[9] CRL and OCLC.
[10] CRL and OCLC, Appendix 3: Minimum Required Documents, p. 81.

acquisition of content (ingest); information and access management; and systems infrastructure, technologies, and security.[11]

OAIS and InterPARES Compliance

H-Net currently complies loosely with the six major high-level responsibilities of an archive, as defined by the OAIS reference model. [12] These include:

- **Negotiates for and accepts information.** Guidelines for the types of information acceptable for posting, accepted formats (plain text, with no attachments), and policies for dispute resolution and self governance are provided in the H-Net constitution and by-laws.[13]

- **Obtains sufficient control for preservation.** Copyright is retained by message authors with H-Net reserving permission for electronic distribution rights, as noted in the section "How H-Net Works."

- **Determines designated consumer community.** H-Net's designated community consists of its editors and contributing scholars, as well as the interested general public.

- **Ensures information independently understandable.** Message headers contain metadata for users to determine context and provenance, including basic subject area of coverage (list), subject, author, and date.

- **Follows established preservation policies and procedures.** Current preservation activities include MATRIX's backup processes and security measures, as described above. The creation of MD5 hashes for each message provides an informal means for ensuring authenticity when a researcher attempts to access them; if a bad URL webpage appears instead, authenticity has been compromised.

- **Makes the information available.** Researchers may browse and search for information in H-Net through a web-based interface, as described in the section "How H-Net Works."

[11] See http://www.h-net.org/archive/documentation/TRAC%20current%20publish.pdf for the original TRAC assessment of the H-Net preservation system.
[12] CCSDS, 3-1 – 3-5.
[13] http://www.h-net.org/about.

Components of the message ingest, storage, and retrieval processes in the H-Net preservation system map to the Information Packages (IPs) defined by the OAIS model.[14] The Submission Information Package (SIP) corresponds to the message, including the body and some header information, posted by an editor. The Archival Information Package (AIP) corresponds to the message as well, and can include the cached metadata used for more expeditious message retrieval. A selected message displayed in the browser window or brought up using LISTSERV commands, plus a subset of metadata pulled from the notebook file, corresponds to the Dissemination Information Package (DIP). As collections of messages, the notebook files may be considered special cases of AIPs known as Archival Information Collections (AICs).[15] (See Figure 4.)

The H-Net preservation system does not currently comply with the InterPARES guidelines for ensuring authenticity, which require that the preserver can demonstrate that records retain their integrity throughout an archival repository's ingest, maintenance, and dissemination processes.[16] Editors and authors can check the content of posted messages, and a retrieval attempt may turn up a bad URL if a message has been compromised. But the MD5 hashes calculated for messages and used for discovery purposes are not employed for fixity checks, a practice for ensuring the authenticity and integrity of electronic records. Likewise, no fixity measures are in place for the collections of messages, the notebook files. The lag time between when an editor sends a message for posting and when it is actually assigned an MD5 hash poses the biggest obstacle to ensuring authenticity.

[14] CCSDS, 2-5 – 2-7.
[15] CCSDS, 4-37 – 4-39, 4-42 – 4-43.
[16] International Research on Permanent Authentic Records in Electronic Systems (InterPARES), *The Long-Term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, "Appendix 2: Requirements for Assessing and Maintaining the Authenticity of Electronic Records, March 2002, http://www.interpares.org/book/interpares_book_k_app02.pdf, 7.

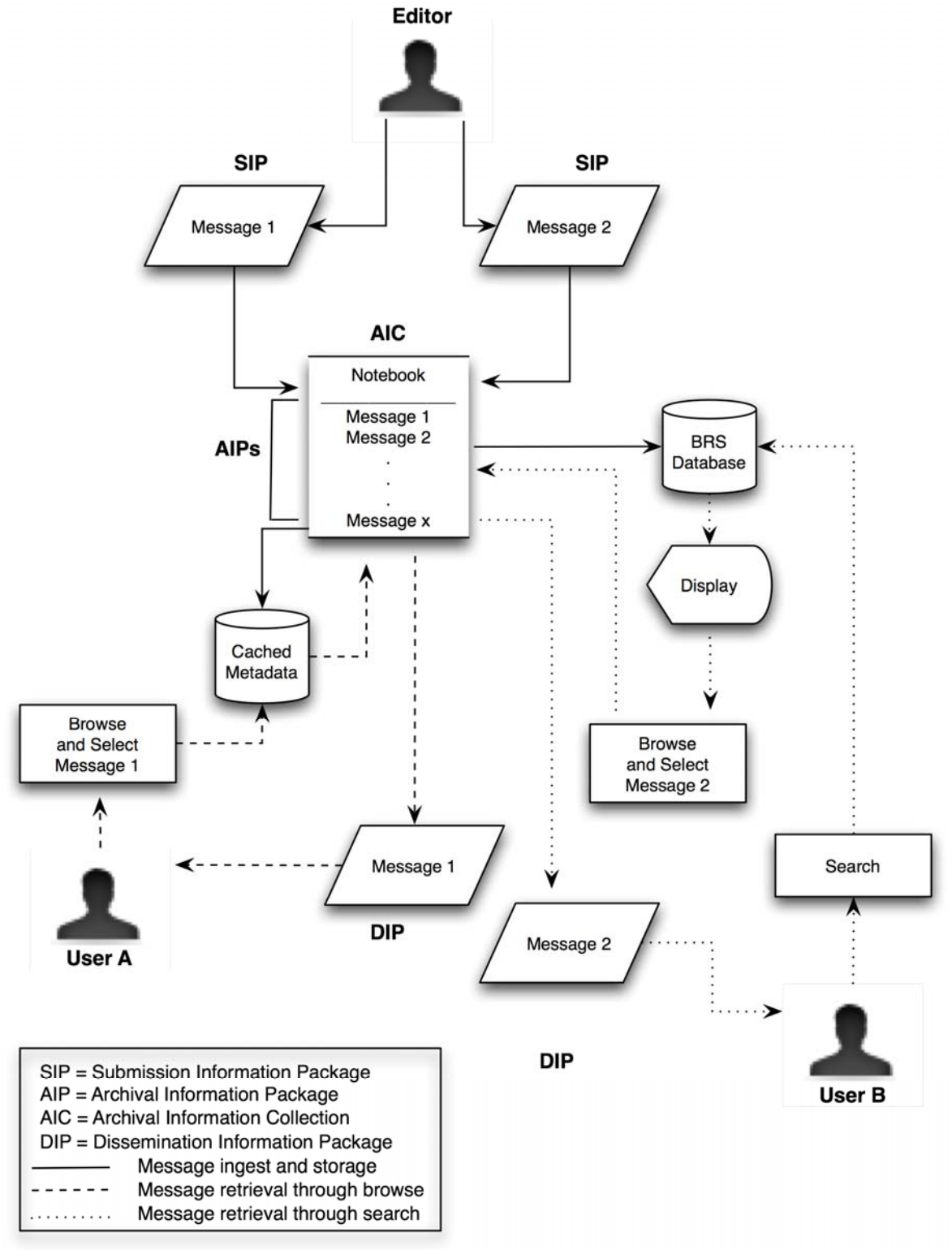

Figure 4. H-Net Message Ingest, Storage, and Retrieval Processes

Analysis

Consultations on the preservation system assessment with an archival advisory board, as well as with H-Net staff and council members, resulted in the following improvement plans.

Backup and Storage

Several improvements must be made to MATRIX backup and storage processes to better protect and ensure continued availability of H-Net and other data. To begin with, the systems administrator must either install a lock on the cabinet that houses the permanent backup tapes or relocate the tapes to more secure local storage.

In addition to maintaining the "permanent" backup tapes, MATRIX is planning a reciprocal storage arrangement with the Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan, Ann Arbor. ICPSR will synchronize and copy over MATRIX data into dark storage, and MATRIX will do the same for ICPSR. MATRIX is also exploring the option of creating a second set of permanent backup tapes and storing them offsite at a secure, temperature controlled storage facility in nearby Lansing, Michigan. This will likely be accomplished through an arrangement with the Michigan State University Archives, which already contracts with the facility. Rather than "permanent," all of these tapes will be considered long-term backup copies and placed on a two-to-three year retention schedule.

Archival Storage

MATRIX must implement a true archival storage plan for H-Net. On an annual basis, MATRIX will copy the H-Net records and associated metadata created during that year onto tapes, along with a text file containing provenance information for the archival copy. This provenance metadata will consist of information about when, where, and on what type of media the archival copy was made, as well as metadata related to such actions as media refreshment as they take place in the future. One copy of these tapes should be kept at an offsite location, such as the aforementioned storage center in Lansing, and a second copy should be kept in a secure location on the MATRIX premises or elsewhere, with media refreshment scheduled for every five years. MATRIX will keep a wiki-based log, similar to that established for the backup tapes, containing descriptive and provenance metadata for each tape and the actions taken on the data.

While MATRIX is committed to maintaining and preserving the H-Net archive, the center is taking the prudent measure of identifying a possible successor in the archive's stewardship. The ideal potential partner would also provide an alternative archival storage repository for H-Net even as MATRIX continues to function as the live host, holding the records and metadata in a dark archive. MATRIX could provide the partner

with a current copy of the H-Net data and associated metadata, along with new data for each successive year on an annual basis.

MATRIX should also strive to participate in a distributed storage system. Within the next two to five years, Michigan State University plans to implement the Integrated Rule-Oriented Data System (iRODS)[17] or a similar rules-based preservation system that MATRIX could join. Another option might be participation in a Lots of Copies Keep Stuff Safe (LOCKSS)[18]- or Storage Resource Broker (SRB)[19]-based system.

Authenticity

As MATRIX suspected before embarking on its assessment of the H-Net preservation system, H-Net has many issues with ensuring the authenticity of messages and associated metadata. These include the lack of fixity checks for messages and notebook files; the changing of metadata to reflect the editor's information rather than that of the original author; and system loopholes that allow editors to delete or make changes to notebooks.

Fixity. To ensure authenticity of messages and notebook files, fixity must be established and checked periodically using message digest algorithms. For messages, the system must assign MD5 hashes on posting rather than waiting up to several days. The MATRIX systems administrator has approached L-Soft about changing the LISTSERV software to enable immediate MD5 hash assignment. As this change in functionality could take months or even years to implement, the administrator will take the stopgap measure of making programming modifications that will ensure the assignment of an MD5 hash to a message within 24 hours of posting. Message digest calculations would be performed on all of the messages posted during the seven-day period before the notebook containing those messages was created. At time of notebook file creation, hashes will also be created for the notebooks. Message digest calculations will then be performed on a weekly basis to ensure notebook file integrity. Any errors in message digest calculation will be logged and manually investigated.

The additions of fixity checks will change how the H-Net preservation system maps to the OAIS model. Fixity databases for the messages and notebooks will provide fixity information as part of each AIP and AIC. (See Figure 5.)

[17] http://www.irods.org/index.php/Introduction_to_iRODS
[18] http://www.lockss.org/lockss/Home
[19] http://www.sdsc.edu/srb/index.php/Main_Page

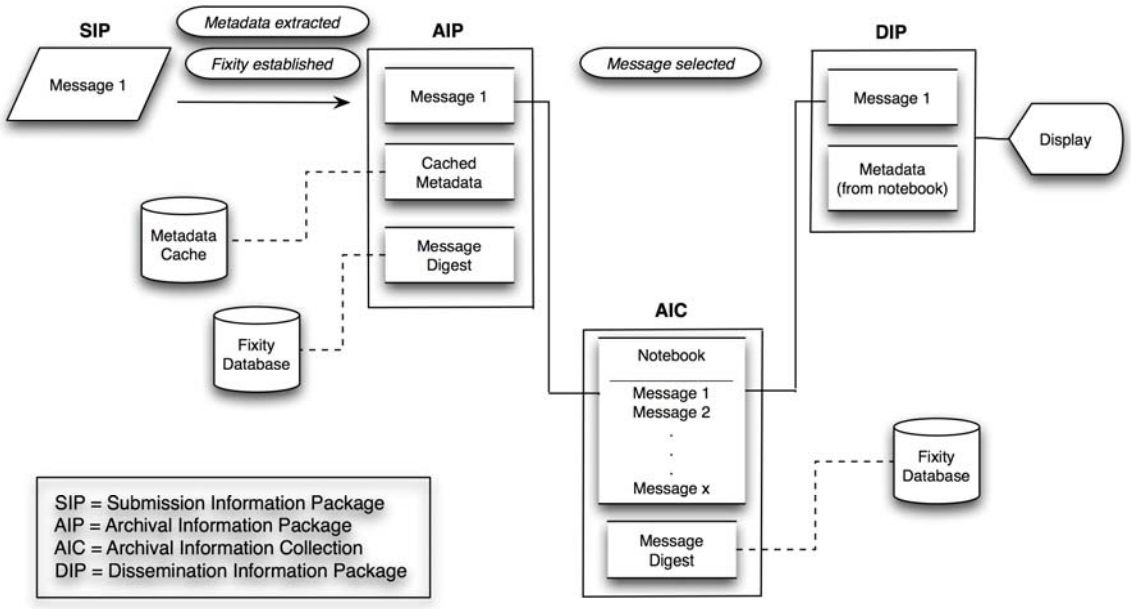

Figure 5. H-Net Information Packages, with Fixity Information

Consideration is also being given to running digital signatures on list catalogs for each H-Net list. This digital signature would be updated when a new notebook posts. Periodic checks would ensure against the deletion of notebook files.

Accurate message creation metadata. To correct the problem of inaccurate message creation metadata that arises when an editor makes a change to a message before posting, a web-based list editing interface had been proposed. This interface would ensure the maintenance of provenance information by automatically retaining the original metadata along with that of the editor. Legacy messages would not benefit from this improvement, however, and the H-Net Council decided against expending development resources on the new interface.

Restriction of administrative capabilities. To eliminate a loophole that allows editors to delete or make changes to notebook files, notebook rights will be restricted to MATRIX and H-Net staff with postmaster privileges. Staff with access to root accounts will retain those privileges, however. The likelihood of staff tampering with H-Net files is low, and the need to ensure 24/7 availability of MATRIX systems—including online history courses hosted by MATRIX—is too important.

File Formats and Preservation

Most messages and notebook files are in plain text formats such as ASCII, Unicode, and UTF-8. Attachments to messages on the private lists are in proprietary formats, including various versions of Microsoft Office applications and PDF.

Messages and notebooks. The plain text formats of messages and notebook files are non-proprietary, well-documented, recommended archival formats for text that require no migration strategy at this time.[20]

Migration strategy for attachments. As the attachments are in proprietary formats at risk of obsolescence, they do require a migration strategy. The systems administrator will first conduct an inventory of attachment formats. Conversion tools for the most commonly occurring file formats will then be provided, or the user will be pointed to a website containing such conversion tools. To keep up with the development of new formats, a technology watch will be established or leveraged. Consideration was given to normalizing attachments to open source formats on ingest, thus minimizing or eliminating the need for format conversion on retrieval. As fewer than two thousand of the one million H-Net messages contain attachments of documents and other files of interest, however, MATRIX decided not to normalize attachments at this time.

Browser access to private lists. To enable access to messages and attachments, the private lists must be made browsable in the manner of the public lists. The systems administrator will enhance the existing software to provide constructed URLs and the browser interface for private list messages. In order to browse the private lists, users will still require appropriate permissions and authentication.

Other Possible Technical Improvements

Other improvements that MATRIX is considering to the H-Net preservation system include preserving links within messages to their original content and creating shorter persistent message URLs.

Preservation of links to original content. To ensure continued access to web links within messages, the systems administrator will explore methods of redirecting URLs to archived websites in case the original site gets taken down. This will most likely be accomplished through redirects to the Wayback Machine of the Internet Archive,[21] which periodically sweeps the indexable Web to capture websites at a given point in time. Although these redirects might not catch every mothballed website, they would capture many or most of them.

Shorter persistent URLs. The URLs that identify unique messages are long and cumbersome for use citations. To make the URLs more user-friendly for bookmarking and citations, MATRIX may develop a system of mapping the actual long URLs to shorter ones.

---

[20] Lee, Bronwyn, Gerard Clifton and Somaya Langley, "PREMIS Requirement Statement Project Report," Appendix 2: Recommended list of supported formats, p. 25, Australian Partnership for Sustainable Repositories (APSR), National Library of Australia, July 2006, http://www.apsr.edu.au/publications/presta.pdf.

[21] http://www.archive.org

Problems Uncovered by TRAC

As noted in the "Case Methodology" section, use of the TRAC to assess H-Net as a preservation system showed adequate compliance with most sets of criteria. The gap analysis revealed that there is room for improvement, however. For example, although the H-Net Strategic Plan includes a reference to commitment to "permanent archiving" of content,[22] there should be a more detailed, explicit commitment to preservation as part of the system's Mission Statement, as required by criterion A1.1, "Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information."[23] Other criteria stipulate requirements for documenting the technology history of the repository and policies regarding staff roles and responsibilities. The archivist must document and in some cases create policies explicitly documenting how H-Net meets the minimum requirements of a trusted digital repository and other appropriate criterion.

Some areas of the TRAC currently are not sufficiently addressed by, or documented for, the H-Net system. For one, a succession plan has not been put into place, as stipulated by criterion A1.2.[24] MATRIX is currently negotiating with possible partner institutions regarding a succession plan for H-Net, and a statement indicating a commitment to identifying a partner with specific capabilities will fulfill the requirement. Other areas that must be addressed and adequately documented include creation of the archivable package; preservation planning; archival storage; and preservation/maintenance of AIPs.

TRAC as Evaluation Tool

TRAC proved to be a thorough yet flexible means of assessing the H-Net archive's viability as a trusted digital repository. Each criterion left room for interpretation, and each provided several options for supporting evidence and documentation. Running through the 84 criteria was somewhat tedious, although it required only approximately one week of rigorous consultations with the MATRIX systems administrator and office manager as well as the associate director of H-Net. (Note that this short assessment period belied the fact that the archivist spent more than two months conducting research on H-Net and digital preservation strategies. Also, H-Net is a relatively small and homogenous data set created mostly in recommended, non-proprietary formats for text. Larger, more complex digital repositories with greater numbers of administrators would likely require more time for a TRAC audit.)

The process resulted in a fair snapshot of the current state of the H-Net archive, clarifying what's needed to "narrow the gap" between the current and desired states of the H-Net preservation system. After implementing the improvements to the system noted earlier in this section and creating appropriate supporting policy documentation, a new TRAC

[22] H-Net: Humanities and Social Sciences Online, "Strategic Plan 2005," http://www.h-net.org/about/strategic.php. Retrieved 21 November 2008.
[23] CRL and OCLC, p. 10.
[24] Ibid.

assessment will be performed. Commitment to regular audits is a TRAC criterion (A3.9[25]), and will be especially important for repositories seeking certification for use by third-party depositors. Periodic internal audits, identification of areas requiring improvements, and the implementation of those improvements will bring MATRIX closer to establishing H-Net as a trusted digital repository.

The H-Net archive contains a relatively small, homogenous set of data that nonetheless requires a systematic preservation strategy. Running a TRAC audit on H-Net and the subsequent development and implementation of a repository improvement plan illuminated the benefits the use of TRAC could offer to more complex repositories. Later this year, Michigan State University's office of Libraries, Computing & Technology will leverage MATRIX's experience using TRAC on the H-Net archive in its design of an institutional repository for the university.

[25] CRL and OCLC, p. 15.