

Preservation of the H-Net E-Mail Lists: Suggested Improvements

Lisa M. Schmidt
MATRIX: Center for Humane Arts, Letters and Social Sciences Online
Michigan State University
August 2008

Preservation of the H-Net E-Mail Lists: Suggested Improvements

Michigan State University, through its research and outreach unit MATRIX: Center for Humane Arts, Letters and Social Sciences Online in partnership with H-Net: Humanities and Social Sciences Online, seeks to advance the state of electronic mail preservation by assessing and improving digital preservation practices for the H-Net electronic mailing list. H-Net, an international consortium of teachers and scholars, maintains the largest and oldest online collection of born-digital, contemporaneously generated, and content moderated arts, humanities and social science material on the Internet, including more than 180 free interactive lists containing more than one million e-mail messages. In addition to these public lists, H-Net includes more than 230 “private” lists used by editors, board members and administrators for planning, testing, and advisory purposes.

MATRIX’s electronic records archivist has assessed existing H-Net preservation practices, and the results of that assessment may be found in the document “Preservation of the H-Net E-Mail Lists: Current Practices”.¹ In keeping with the InterPARES guidelines for assuring the authenticity of H-Net records and the NARA/OCLC Trusted Repositories Audit & Certification: Criteria and Checklist (TRAC), and with the guidance of an archival advisory board, several areas requiring improvement were identified.

¹ Schmidt, Lisa M., “Preservation of the H-Net E-Mail Lists: Current Practices,” MATRIX: Center for Humane Arts, Letters and Social Sciences Online, Michigan State University, March 2008, <http://www.h-net.org/archive/documentation/H-Net%20Current%20Practices%20Post2.pdf>.

With the MATRIX systems administrator, the archivist has analyzed the problem areas and developed a set of suggested technical measures that may be taken to improve the preservation system. The technical improvements that will be described in this document include measures to better ensure authenticity, to preserve attachments on the private H-Net lists, to preserve links to content, and to improve usability.

Note that suggestions for improvements to the major areas of backup and storage are not included here. That discussion will be forthcoming following an analysis of current best practices in records management.

In addition to managing the implementation of the technical measures that will improve the H-Net archive preservation system, the archivist will again vet H-Net against the TRAC. Support for the TRAC criteria will be more comprehensively substantiated and documentation of policies will be created as needed. This will likely include and will not be limited to the documenting of H-Net's technology history; policies delineating staff roles, responsibilities, and authorizations; change management policies; and a disaster recovery plan.

The remainder of this document will outline the suggested technical measures for improving the H-Net preservation system. For more in-depth explanations of the H-Net system and current preservation practices referenced here, please see the document "Preservation of the H-Net E-Mail Lists: Current Practices."²

² Ibid.

To Better Ensure Authenticity

A major goal of the H-Net e-mail list preservation project is to “ensure preservation, authenticity, and sustainability” of the information in the lists. While it is impossible to guarantee authenticity of electronic records with absolute certainty, several measures can be taken that would mitigate risks to the authenticity of H-Net messages.

Fixity

The nature of digital information technology, including the ability to change or make copies of a file, leaves an electronic record open to inadvertent or intentional altering even after its submission into an archive. The guidelines of the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) thus indicate: “Determining that [an electronic record] is free from tampering or corruption means demonstrating that its integrity remains intact through space and time.”³

One means of demonstrating that integrity is assigning a message digest, or cryptographic hash function, to the electronic record on submission into a digital archive. Periodic message digest calculations on the record will ensure that it has remained authentic or alert administrators that something has happened to change it.

The H-Net lists currently use the MD5 algorithm to calculate message digests that then identify individual messages for discovery and citation purposes. To date, however, no message digests have been used to actually establish fixity of the messages as digital objects are submitted and stored in H-Net notebook files.

³ International Research on Permanent Authentic Records in Electronic Systems (InterPARES), *The Long-Term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, “Appendix 6: How to Preserve Authentic Electronic Records,” 2001, http://www.interpares.org/book/interpares_book_o_app06.pdf, 3.

H-Net messages are submitted as individual files; in Open Archival Information System (OAIS) terminology, they are the submission information packages (SIPs) in this preservation system. Seven days worth of messages in any given list are grouped together and stored in “notebook files”; these function as the archival information packages (AIPs) in the system—the actual H-Net archive. To better ensure authenticity, message digests must be calculated for both the individual messages on submission and the notebooks once they are completed.

Individual Messages

The following measures must be undertaken to better ensure the authenticity of individual messages:

- **LISTSERV creates MD5 hash for each message at time of submission.** Messages submitted to the H-Net lists currently receive hashes created using the MD5 algorithm, as noted above. Due to the way LISERV currently functions, however, there can be a lag time of up to seven days after submission before a message is actually assigned a hash. This lag time essentially invalidates any claim to authenticity of the H-Net messages. To ensure the authenticity of the messages in the H-Net archive, they must receive hashes on submission.

Two actions are being proposed to correct this problem. First, LISERV vendor L-Soft has been asked if it could rewrite future versions of the software to allow this capability. L-Soft has responded favorably to the request; however, it could take several months or even years for the change to be made. Therefore, a second action—developing a process to calculate hashes within 24 hours of message submission—is being proposed. The calculations for a day’s worth of messages could, for instance, occur at midnight every day. This should be an acceptable time window, and is certainly an improvement over the current situation.

- **Message digests and other currently cached metadata continue to be sent to log browse cache.** The message digests calculated on submission or in the smaller time window, per the above, will continue to be added to the log browse cache for message identification and discovery purposes.
- **Message digests also sent to separate database for use in fixity checks.** In addition to being sent to the log browse cache, the message digests should be sent to a separate database set up solely for use in fixity checks.
- **Hashes validated on completion of notebook file.** An automated checksum process must be developed, to occur on completion of a notebook file. If the checksums match, the notebook file itself will be ready for archiving and subjected to its own message digest calculation. (See next section.) If the checksums do not match, an alert will be generated. Any errors will be corrected through manual intervention, most likely by checking against an earlier backup of the system.

Notebook Files

The following measures must be undertaken to better ensure the authenticity of notebook files:

- **SHA-2 message digest created for notebook on completion, after checksums successfully run on messages; sent to database for use in fixity checks.** A system must be developed for calculating SHA-2 message digests for notebooks in which the individual messages have passed their checksum tests. SHA-2 should be used rather than MD5, as the integrity of MD5 has been compromised and use of SHA-2 is currently considered a best practice. (Note that the decision was made to continue to use MD5 to calculate message digests for individual messages because: (a) the current system works well for message identification and discovery purposes; and (b) the likelihood of a compromising collision occurring during the “live” seven day period is very low.)

The SHA-2 message digests will be sent to a database created to store them. SHA-2 message digests will also be calculated for all of the existing notebook files, with the understanding that the individual messages within them were never subjected to fixity checks.

- **Hashes validated on a regular periodic basis.** An automated checksum process must be developed for the notebook files, to occur on a weekly basis. If the checksums do not match, an alert will be generated. Any errors will be corrected through manual intervention, most likely by checking against an earlier backup of the system.

Accurate Message Creation Metadata

In the current system, a message that an editor modifies essentially becomes a new message. LISTSERV has no provision to maintain the original creation metadata; the author and date written (creation date) metadata thus changes to reflect the editor's name and the current date. The editor can manually enter the original message's author, date, and subject. This method of changing the metadata is labor intensive, prone to error, and not done with nearly one-third of the lists, however. By not automatically maintaining original metadata alongside the editorial modification metadata, the authenticity of the message metadata is compromised.

Note that this lack of accurate creation metadata can also compromise discovery of a message, as researchers cannot view a comprehensive list of messages sorted by author names. Messages by particular authors may still be found using full-text search if the author's name was included within the message or in a signature.

The following measure must be undertaken to better ensure the authenticity of message metadata:

- **Automate addition of editor's metadata to messages, for cases in which editor changes messages.** To automate the addition of an editor's metadata to a message without losing the original creation metadata, a list editing web interface could be built. This could also rectify other problems with messages the editor changes, including the removal of Yahoo! footers, RFC attachments, and Microsoft mailer issues. This interface could only be used with new messages; the metadata for legacy messages would not be improved upon.

Consideration needs to be made as to whether it would be worthwhile for MATRIX to build this interface. Would list editors be amenable to working with this new interface? Is H-Net seriously considering moving to another communication platform, such as a blog format, within the next two years? If so, perhaps building this interface would not be worth the development resources. Creating it would not be a trivial effort, and it could only be used with new messages.

Restriction of Editors' LISTSERV Administration Capabilities

Currently, H-Net list editors have the ability to change and even delete entire notebooks from the H-Net file system using e-mail commands. Changes to a notebook would result in checksum errors that require manual intervention, putting an unnecessary squeeze on staff resources. Deletions of entire notebooks could occur without being detected. Although there is no record to date that any editor has ever edited or deleted a notebook file, the existence of this loophole clearly compromises the authenticity of the H-Net archive.

The following measure must be undertaken to better ensure the authenticity of the notebook files against possible editor action:

- **Eliminate editors' ability to retrieve and change notebook files.** Notebook rights must be restricted to MATRIX postmasters. Currently, these postmasters include the

systems administrator, the executive and associate directors of H-Net, and the director and associate directors of MATRIX.

Possible Security Threats to Authenticity

In the current LISTSERV setup, anyone at MATRIX who has organization-wide system account privileges could potentially tamper with the H-Net messages. MATRIX staff members with these privileges include the systems administrator, three lead programmers, and the director and associate directors of MATRIX. After much discussion over the possibilities of limiting access permissions to LISTSERV and/or developing the means to generate audit logs of non-standard activity, a risk analysis resulted in the decision to not implement either change.

Several people must hold system account privileges to ensure 24/7 availability of all of the systems hosted by MATRIX. In addition to H-Net and the digital libraries hosted by MATRIX, this includes online history courses that students must have ongoing access to for fulfillment of course requirements. Restricting root account access to LISTSERV administration would be difficult if not impossible. Likewise, developing the means to generate audit logs of non-standard activity would be difficult, time-consuming, and probably not of much benefit anyway.

Valuable scholarly resource that H-Net is, the nature of the information contained in the H-Net lists does not merit the highest levels of security usually reserved for classified government documents, records subject to legal regulation, and records containing sensitive personal information. Those who hold root account privileges at MATRIX are trusted employees. The risk of any of them deliberately compromising the H-Net lists or any of the systems hosted by MATRIX is minimal to nonexistent. A

hardware or software failure is more likely to occur, and that risk can be mitigated by MATRIX's backup system.

To Preserve Attachments on Private Lists

The public H-Net lists require plaintext postings and do not allow attachments. Currently, there is no need for a migration strategy for these discussion logs as the open ASCII and UTF-8 standards are readily available and content stored in the logs may be accessed using text viewers. Posting requirements are not so strict for the private lists, however, which do include HTML-formatted messages as well as attachments in proprietary formats.

Used primarily for administrative rather than scholarly purposes, the private H-Net lists are of secondary importance in the preservation effort. The private lists may be of some historical interest, however, and should receive preservation consideration.

To that end, a migration strategy is required for the attachments to the messages on the private lists. The private lists must also be made browsable so that users may access the attachments in the first place.

Browser Access for Private Lists

The following measures must be undertaken to provide browser access to messages and attachments on the private lists:

- **Private lists made browsable.** Currently, only a subset of the private lists are browsable in the manner of the public lists. Metadata for private list messages must be stored in the log browse cache and used to construct URLs for browsing purposes.

Permissions and/or authentication may still be required to browse, search for, and view messages in these lists. The existing software must thus be enhanced to ease access to the private lists and to enable access control.

- **Attachments made available through browser.** A download link to an attachment will be provided with its respective message. Attachments will be made available for download and viewing in their native formats as part of the retrieval (dissemination) process. If a message is available in alternative formats, such as HTML, a link to that may be provided as well.

Note that attachments will not be detached and stored separately. Due to the nature of e-mail, they are actually embedded in the messages rather than “attached” to them. Taking messages apart and storing the pieces separately would add a level of complexity to the preservation system that could unnecessarily compromise the integrity of both messages and “attachments.”

Migration Strategy for Attachments

The following measures must be undertaken to provide a migration strategy for attachments on the private lists:

- **Inventory/audit of file types, number of each file type.** A full inventory of the attachments on the private, H-Net-related lists must be conducted in order to determine which formats to migrate. This will require identification of relevant lists by MATRIX and H-Net management.
- **Conversion on demand.** A technology watch will be established or leveraged to alert the H-Net systems administrator of any updates or changes to the formats of attachments in the system. As new versions of software come available, it will become necessary to provide the means for users to convert a downloaded attachment to the latest version. File format conversion tools could be kept in reserve on the H-Net website; at the very least, links could be provided to other websites that offer such tools. Alternatively, MATRIX could provide the means for automatic conversion to the

new file formats. The resources required to enable and maintain this service as well as the low expected demand for it render the latter option less worthwhile.

In either case, it will be necessary to limit conversion resources to the most popular formats—such as Microsoft Office— and/or those with readily available conversion tools. It may be impossible or not worth the effort to convert files in obscure, little used formats or formats for which conversion tools are not available. Also, conversion can be made available only for attachments for which the correct MIME types have been provided.

Note that the new files created by or for the user during the conversion process will not be added to the H-Net preservation system. As much as possible, conversions will be from the original files.

To Preserve Links to Content

Some newer messages on the H-Net Lists include URLs that link to articles and websites of interest. Due to the dynamic nature of the Internet, there is a chance that over time these links could break. To preserve the messages so that the content matches the original as closely as possible, it will be necessary to develop a means to maintain access to the contents of those links.

URLs Within Messages Linked to Archived Web Pages

The following measure must be undertaken to preserve links to original content of URLs within messages:

- **Maintain links to original content.** “Broken” URLs within H-Net messages will be identified and redirected to the original websites archived in the Wayback Machine of the Internet Archive (www.archive.org). As the Wayback Machine is not infallible, it is

understood that links to every bit of original content may not be possible. However, this method should work in most cases.

To Improve Usability

Currently, the log browse application builds URLs for each posted message when the month view is generated from the Discussion Logs section of a given list. Each URL incorporates a combination of a message's filename and MD5 hash, providing each message with a unique, persistent name for citation in journals and other published materials. These URLs are very long and while logically constructed, are not easily read or easy to key. Shorter persistent URLs would be more inviting for use in citations—particularly in printed materials— and less prone to input error.

Note that this is not, strictly speaking, a preservation activity. Shorter, less cumbersome URLs would make for a “cleaner,” neater user experience, however.

Shorter Persistent URLs

The following measure must be undertaken to provide users with shorter persistent URLs:

- **Creation and mapping of shorter, more user-friendly message URLs.** A naming scheme for the shorter URLs would need to be developed and approved by H-Net management, the council, and the editorial board. Creation of the new URLs could be automated and mapped to the actual URLs. As this would only be practical for new messages, consideration must be made as to whether it would be worth the work involved—particularly if H-Net is moving to a blog or other Web 2.0 communication platform within the next two years.