H-Net:
Preserving and Improving Access to Specialized
Electronic Mailing List Archives

Final Narrative Report to NHPRC

Prepared by Lisa M. Schmidt, Electronic Records Archivist
MATRIX: Center for Humane Arts, Letters, & Social Sciences Online
Michigan State University
February 2010
Executive Summary

H-Net: Humanities and Social Sciences Online, an international consortium of scholars and teachers hosted by the MATRIX digital humanities research center at Michigan State University, is the oldest collection of born-digital, content-moderated arts, humanities, and social science material on the Internet. At the heart of H-Net lies more than 180 public academic networks, or e-mail lists, covering a wide range of humanities and social science topics, as well as more than 230 private administrative lists. H-Net list messages have grown to number more than one million since H-Net began in 1992, and the H-Net list archive is considered a valuable record of academic discourse. In 2007, the National Historical Publications and Records Commission (NHPRC) awarded MATRIX with a grant to conduct research on the preservation and access functions of the H-Net e-mail list archive that would ensure longevity of the content.

The first and primary goal of the project, ensuring the preservation of the H-Net list archive, began with the documentation and analysis of MATRIX backup practices and H-Net list processes. H-Net runs on LISTSERV e-mail list software, and the LISTSERV message distribution workflow as well as a web-based public list interface and other functions developed by MATRIX allowed the H-Net list archive to map adequately to the Open Archival Information System (OAIS) reference model. Message content and metadata are created in archival plain text file formats, so no format migration strategy was required. The existing system failed to comply with the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) authenticity guidelines, however. Also, an evaluation of the H-Net list archive against the Center for Research Libraries (CRL) and Online Computer Library Center (OCLC) Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist revealed large gaps between the then-current state of the archive and a trusted digital repository.

A preservation improvement plan called for better authenticity measures, especially establishing and regularly checking fixity of messages and “notebook” files containing collections of messages. Other major preservation enhancements included a plan for preserving attachments to messages on the private lists, many of which were in proprietary file formats; offsite backup plans for MATRIX, including a reciprocal storage arrangement with the Inter-University Consortium for Political and Social Research (ICPSR) and a plan to store backup tapes away from the MATRIX premises; and the decision to document digital preservation processes and procedures, as TRAC emphasizes the importance of written documentation as evidence of meeting criteria.
After review with the project’s archival advisory board and approval from the H-Net Council editorial board, most of the recommended preservation improvements were implemented. Digital preservation policies for the H-Net list archive were documented using a digital preservation management methodology developed by researchers at Cornell University and ICPSR. Long-term preservation needs were also indicated, including commitment to ongoing TRAC evaluations and periodic digital preservation policy reviews; participation in university-wide and regional distributed preservation systems; and the formalization of a succession plan for the H-Net list archive in the event that MATRIX can no longer provide the hosting service.

The secondary goal of the project, enhancing access to H-Net through improved search tools, involved research into Semantic Augmented Consensus Clustering (SACC) techniques. This research met with limited, mixed success, and the clustering techniques had not gone beyond the prototype stage and were not tested on H-Net e-mail records before funding was depleted.

During the course of the project, information about its progress was disseminated through a project website, published articles, and presentations. The project website, http://www.h-net.org/archive, went live during the first half of 2007. Articles were developed for online publications and key journals, including American Archivist, and presentations were delivered at Society of American Archivists (SAA), Midwest Archives Conference (MAC), and other strategic conferences and meetings.

In conclusion, the project met the goal of ensuring authenticity, preservation, and persistence of the H-Net e-mail list archive, but the research in improving access through clustered search techniques fell short of producing useful results. The TRAC and other tools and methodologies used to evaluate and improve on the preservation of the H-Net e-mail lists may be applied to other e-mail list archives as well as other large sets of data. In fact, they are being used to analyze and develop preservation plans for other repositories of electronic records at MATRIX and Michigan State University.

Mark Kornbluh
Principal Investigator

Lisa M. Schmidt
Project Manager
1. Introduction

At the heart of the international scholarly consortium H-Net: Humanities and Social Sciences Online lies a set of e-mail lists containing many years of academic discourse on humanities and social sciences topics. In 2007, MATRIX: Center for Humane Arts, Letters, & Social Sciences Online, a digital humanities research center at Michigan State University that hosts H-Net, received a grant from the National Historical Publications and Records Commission (NHPRC) to conduct research on the preservation and access functions of the H-Net e-mail list archive that would ensure longevity of its content.

1.1. H-Net and the H-Net E-Mail Lists

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 The oldest collection of born-digital, content-moderated arts, humanities, and social science material on the Internet, H-Net began in 1992 as a virtual service hosted at the University of Illinois at Chicago. Since then, H-Net has grown to include more than 180 scholarly social sciences and humanities networks hosted by MATRIX: Center for Humane Arts, Letters, & Social Sciences Online, a digital humanities research center at Michigan State University. Content includes *H-Net Reviews*, the largest online scholarly book review journal in the world; job and meeting announcements; and 185 free, public academic networks, or electronic mail (e-mail) lists. A 14-member council governs H-Net’s policies and activities.

The academic e-mail lists at the heart of the H-Net consortium cover a wide range of humanities and social science topics. More than 450 editors and 130,000 subscribers participate in the networks, and postings average 5,000 per month. An estimated 230,000 messages were viewed during the last week of April 2009 alone. In addition to its public lists, H-Net includes more than 230 “private” lists used by editors, board members, and administrators for planning, testing, and advisory purposes. There are more than 1 million e-mail messages in the H-Net list archive, and that number continues to grow.

---

1.2 MATRIX: Center for Human Arts, Letters, & Social Sciences Online

MATRIX, the digital humanities research center that hosts H-Net, is devoted to the application of new technologies in teaching, research, and outreach. The center creates and maintains digital libraries of humanities and social science materials, provides training in computing and new teaching technologies, and creates forums for the exchange of ideas and expertise. Along with the H-Net scholarly community, MATRIX hosts the major digital library repositories African Online Digital Library (AODL), Detroit Public Television’s American Black Journal video archives, Historical Voices, and The Quilt Index, among others.

1.3. H-Net Preservation and Access Research Project

In 2007, MATRIX received a grant from the National Historical Publications and Records Commission (NHPRC) to advance the state of electronic mail archiving by assessing and improving upon the digital preservation, management, and access functions of the H-Net e-mail lists. Representing a compilation of years of academic discourse, the H-Net lists were considered a valuable scholarly resource worthy of long-term preservation and investigation into more sophisticated techniques for access to the records. Students and practitioners routinely bookmark H-Net messages and cite them in research and publications. Preservation and improved access would ensure future availability of the H-Net lists to readers of scholarly research and could help in providing a deeper understanding of the context and evolution of the represented fields. As a study of preservation and access for a born-digital, “living” archive, it was also thought that this work on the H-Net lists would be useful to archivists and others who manage e-mail lists and other large collections of electronic records.

To that end, the two primary goals of the project were:

1. Ensure the preservation, sustainability, and authenticity of the H-Net e-mail lists through use of the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) authenticity guidelines and the Research Libraries Group (RLG) and Online Computer Library Center (OCLC) guidelines on trusted digital repositories.

2. Improve and extend access to the H-Net archive through improved search tools, especially Semantic Augmented Consensus Clustering (SACC) techniques.

In October 2007, MATRIX hired electronic records archivist Lisa M. Schmidt as project manager for the H-Net e-mail list preservation portion of the project. Dennis Boone, H-Net systems administrator, assisted Schmidt with technical analysis of the H-Net list archive and
implementation of recommended technological improvements. Dr. Peter Knupfer, director of H-Net, and Heather Hawley, associate director of H-Net, provided support on administrative issues. The research into search techniques was managed by Bill Punch, Affiliated Faculty at MATRIX and Associate Professor of Computer Science and Engineering at Michigan State University. Dr. Mark Kornbluh, former director of MATRIX and current dean of the College of Arts and Sciences at the University of Kentucky (as of August 2009), was the principal investigator on the project.
2. Preserving H-Net Records

The plan to ensure the preservation, sustainability, and authenticity of the H-Net e-mail lists broke down into the following steps. With minor exceptions and some augmentation, all of the steps below were successfully undertaken and completed.

1. Fully document all current authenticity, preservation, and persistence practices of H-Net.
2. Evaluate these practices in light of the recommendations of the InterPARES framework for preservation of authentic electronic records and RLG’s An Audit Checklist for the Certification of Trusted Digital Repositories.
3. Prepare plan for improvements in H-Net practices for preserving its electronic records—including MD5 hash analysis and potential Lots of Copies Keep Stuff Safe (LOCKSS) participation.
4. Review with project archival advisory board and H-Net governing council and editorial advisory board during steps 1-3 concerning both relevance to H-Net and to large LISTSERV archives more generally.
5. Implement the preservation plan.
6. Formalize a plan for longer-term needs and migration strategy.

2.1. Documentation of Current Practices

Through consultations with Boone, Schmidt fully documented existing authenticity, preservation, and persistence practices. In doing so, she also conducted a thorough analysis of the H-Net ingest, storage, and retrieval processes. As H-Net is hosted by MATRIX, an analysis of the digital humanities research center’s network configuration and backup practices was also necessary. This research resulted in the Preservation of the H-Net E-Mail Lists: Current Practices document, published on the project website in March 2008 and available at http://www.h-net.org/archive/documentation/H-Net%20Current%20Practices%20Post2.pdf. Slightly modified to take into account a physical move MATRIX made soon after the original assessment, those existing practices are outlined below. Note also that some of the references in the original document have changed, such as the definitions of AIPs and SIPs in the H-Net e-mail archive. The terminology used below reflects what was settled on for the duration of the project, as detailed in the supplemental report An In-Depth Look at Information Packages in the H-Net Preservation System, published online at http://www.h-net.org/archive/documentation/H-Net%20IPs.pdf in December 2008.
2.1.1. **Network configuration**

MATRIX runs its operations on several servers. Two of the servers are dedicated to running the LISTSERV software and other applications and processes that allow the H-Net lists to function smoothly. Some processes related to H-Net also share space on other MATRIX servers.

2.1.2. **Backup and security**

MATRIX stores approximately 3 TB of data, including the H-Net e-mail lists, on its servers. The server rack is kept in a climate controlled, physically secured room. (At the time of the original assessment, the servers were located at the MSU Computer Center rather than on the MATRIX premises.) Incremental tape backups are performed daily and a full backup is performed weekly, and those tapes are stored at another building on the MSU campus. As a further safeguard, a full “permanent” tape backup is also performed approximately every four months. These tapes were kept in a cabinet in a minimally secured room on the MATRIX premises at the time of the original assessment—presumably for perpetuity—although no media migration plan was in place.

2.1.3. **How H-Net works**

As H-Net runs on the proprietary LISTSERV e-mail list administration application, many of the processes of the H-Net lists follow the basic functionality of the LISTSERV software. MATRIX has customized some of the basic LISTSERV processes and developed a web-based interface to better serve the needs of the H-Net community.

Message submission policies are such that in most cases, users must be subscribed to a list to post to it. All messages must be written in plain text, and no attachments are allowed on the public lists. When a subscriber sends a message to an H-Net list, the list editor may simply approve the message and send it on for posting or edit it before posting; he or she also has the option to simply ignore the message.

If an editor makes any changes to a message, LISTSERV sees it as a new message and the author and creation date change to reflect the editor’s name and the current date. The editor may take the option of manually entering the original message’s author, date, and subject—a method of changing the metadata that is labor intensive, prone to error, and not done with nearly one-third of the lists. Even if the editor chooses not to enter the message’s original metadata, the author’s name can usually be found in a signature at the end of the message. Researchers can therefore still find messages by particular authors using full-text search, but
they will not see them listed under the original author’s name in a browse list of messages. Figure 1 shows the message posting process.

**Figure 1. H-Net Message Posting Process**

As the editor will only post messages considered part of the academic discourse and thus of current and future interest to researchers, this editorial procedure functions as an appraisal process for the H-Net lists. All messages approved and posted by list editors are permanently archived. At the time of the original analysis, the posting process took from a few seconds up to several days.

Note that some private lists permit subscribers to post messages without going through an editor, in which case no appraisal takes place. The private lists also allow attachments, most in such formats as Microsoft Word (.doc), Microsoft Excel (.xls), Microsoft PowerPoint (.ppt), PDF (.pdf), and JPEG (.jpg).

Approved messages become part of a flat text file, or “notebook.” The H-Net template is set up so that a single notebook includes messages posted to a particular list during a seven-day time period. The messages concatenate in the notebook in the order in which they are received and remain stored in their original order of posting. Figure 2 shows a screen shot of two messages in a notebook file. In terms of description, most of the descriptive metadata for messages is
automatically generated on creation or posting, with the “Author’s Subject” generated by the original author.

**Figure 2. H-Net Notebook File, Showing Messages Posted in Original Order**

| Date: | Mon, 4 Feb 2008 17:47:25 -0500 |
| Reply-To: | H-Net List for British and Irish History <H-A/LBION@H-NE/M.SU.EDU> |
| Sender: | H-Net List for British and Irish History <H-A/LBION@H-NE/M.SU.EDU> |
| From: | "Kelly, Jason" <jaskelly@UIPE.EDU> |
| Subject: | RE: Victorian Attitudes to the Thirty Years War |
| In-Reply-To: | <BLU106-04793CE283E0E0005F0D00S2W3200@bmk.bg> |
| Content-Type: | text/plain; charset="iso-8859-1" |
| Message-ID: | quoted-printable |
| MIME-Version: | 1.0 |
| From: Andrew Connell <andrewconnell@comcast.net> |
| Subject: | RE: Victorian Attitudes to the Thirty Years War |
| Date: | Sun, 2 Feb 2008 19:15:58 -0500 |

> I have a graduate student who is interested in Victorian attitudes
> about the Thirty Years War. Any information that list members can supply
> would be most welcome.

S. R. Gardiner, magisterial historian of England 1603-1656 (18 vols) also wrote ‘The Thirty Years’ War’, published in 1874, described by H. A. L. Fisher as a ‘short, masterly, sketch’. That would be a start.

Andrew Connell

| Date: | Mon, 4 Feb 2008 17:49:01 -0500 |
| Reply-To: | H-Net List for British and Irish History <H-A/LBION@H-NE/M.SU.EDU> |
| Sender: | H-Net List for British and Irish History <H-A/LBION@H-NE/M.SU.EDU> |
| From: | "Kelly, Jason" <jaskelly@UIPE.EDU> |
| Subject: | Final CfP: Beyond camps and forced labour: current international research on survivors of Nazi persecution |
| In-Reply-To: | <04F294930F30664400D7249E8887567D8573885@exchange03.unw.wlv.ac.uk> |
| Content-Type: | text/plain; charset="iso-8859-1" |
| Message-ID: | quoted-printable |
| MIME-Version: | 1.0 |
| From: "Steinert, J.d" <J.D.Steinert@wlv.ac.uk> |
| Subject: | Final CfP: Beyond camps and forced labour: current international research on survivors of Nazi persecution |
| Date: | Sun, 3 Feb 2008 01:10:58 -0500 |

Notebooks are named according to the time periods they cover. Table 1 shows these periods, with days 1–7 of a given month as time period “a,” days 8–14 as time period “b,” and so on. These periods become part of the notebook file name. For example, a notebook with the name “h-africa.log0802a” would be in the H-Africa lists and would include postings from the first seven days (period “a”) of August 2002.
Table 1. H-Net Time Periods

<table>
<thead>
<tr>
<th>Period</th>
<th>Days of Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1-7</td>
</tr>
<tr>
<td>b</td>
<td>8-14</td>
</tr>
<tr>
<td>c</td>
<td>15-21</td>
</tr>
<tr>
<td>d</td>
<td>22-28</td>
</tr>
<tr>
<td>e</td>
<td>29-31</td>
</tr>
</tbody>
</table>

Every 24 hours, the newest messages in the current notebook file are parsed and copied to a proprietary Bibliographic Retrieval Services (BRS) database, where they are available for full-text search. As a completely separate operation, a log browse cache application also reads the notebook messages and creates MD5 hashes for each message seven days after the last message posts to a given notebook. A script then writes message metadata to a MySQL database cache. This metadata includes:

- **filename**—name of notebook file where message is stored, such as h-africa.log0802a
- **offset**—byte position in notebook file where message is stored
- **from**—name and e-mail address
- **subject**
- **dpb**—date posted
- **cbd**—date in a different format for sorting purposes
- **messageid**—MD5 hash

When a user browses a list and clicks on a message to view it, the log browse application pulls the message from the original notebook file, builds a URL for the message, and transforms it into HTML for viewing in the browser. The URL is a combination of the message’s filename and MD5 hash. This serves as a persistent identifier for the message that can be bookmarked for reference and citation purposes. (See Figure 3.)
Figure 3. H-Net Message Retrieval View, with URL Detail

Figure 4 shows how the H-Net ingest, storage, and retrieval processes map to the Open Archival Information System (OAIS) reference model. Messages submitted by the editors are the Submission Information Packages (SIPs). After submission, the messages become Archival Information Packages (AIPs). They are stored in the notebooks, which may be considered Archival Information Collections (AICs). When a user browses and selects a message, the page view received is the Dissemination Information Package (DIP). Alternatively, a user may receive a DIP by searching the BRS database, as it also pulls from the notebooks.
2.1.4. Authenticity, Preservation, and Persistence

In terms of preservation, the backup and storage processes described in the previous section provided one piece of the original practices. There was no means of creating and storing a separate archival copy of H-Net at the time of the original assessment, however, and there was no provision to store a data backup away from the MSU campus.

Authenticity had been based on the author or editor informally checking a message to determine its accuracy. Also, the receipt of a broken URL notification when attempting retrieval might indicate a problem. (Note that H-Net has no record of such a problem ever occurring.) The MD5 hashes generated for the messages were used only for discovery and retrieval purposes, not for...
ensuring authenticity. Even if checksums were calculated, the lag time between when an editor sent a message for posting and when it was actually assigned an MD5 hash would have been an obstacle to ensuring authenticity.

Two administrative loopholes also created opportunities for tampering with H-Net message and notebook authenticity. First, list editors had the ability to change and even delete entire notebooks from the H-Net file system using e-mail commands. Also, several MATRIX employees had organization-wide system root account privileges, and could potentially alter H-Net data.

In terms of ensuring the persistence of the original messages and notebook files, no format normalization or migration strategy was required for the public lists. The most significant property to be preserved is the message content, most of which is in plain text ASCII and UTF-8 formats—and plain text is a recommended non-proprietary, archival format for text. This is in contrast to some well-known preservation projects dealing with proprietary e-mail formats, including the Dutch Digital Preservation Testbed; the Preservation of Electronic Mail Collaboration Initiative among the state archives of North Carolina, Kentucky, and Pennsylvania; and the Collaborative Electronic Records Project (CERP) partnership between the Smithsonian Institution and Rockefeller Archives Center. These projects preserved e-mail in proprietary formats by converting them into XML, another archival format for text, and also by encapsulating messages and metadata in XML for archival storage. In 1999, the Collection-Based Long Term Preservation project of the San Diego Supercomputer Center (SDSC) had likewise experimented with the preservation of Usenet Newsgroup messages using XML.

Unlike the private list messages, some messages on the public lists included attachments in proprietary formats. These would require a migration strategy to ensure future access.

The cached metadata partially fulfills the requirements for Preservation Description Information (PDI) as recommended by the OAIS model. First, the filename provides Reference, Context, and Provenance Information for notebook files. In combination with the filename, an MD5 hash provides Reference Information for an individual message. Additional Context Information for a message may be found in its subject line, and additional Provenance Information may be found in a message’s header. Unfortunately, as described above, key Provenance Information from the message header is lost when the editor makes a change to a message and does not manually add back the creator’s information. At the time of the original assessment, there was no Fixity
Information to help ensure the authenticity of messages and notebook files. Again, the MD5 hashes were not used to establish or verify fixity.

At the time of the original assessment, no documented preservation policies existed for the H-Net e-mail list archive. The Current Practices paper, synopsized here, served as the only such documentation.

2.2. H-Net Preservation, InterPARES, and the Trusted Digital Repositories Checklist

The “authenticity” measures employed by the H-Net e-mail list archive at the time of the original assessment fell well short of the InterPARES recommendations. According to the InterPARES guidelines: “An authentic record is one that is what it purports to be and that is free from tampering or corruption. Determining that it is what it purports to be means confirming its identity. Determining that it is free from tampering or corruption means demonstrating that its integrity remains intact through space and time.”

Without the means to perform fixity checks to ensure integrity and authenticity, existing H-Net preservation practices violated the InterPARES baseline (minimum) requirements supporting the production of authentic electronic records, Requirement B.1: Controls Over Records Transfer, Maintenance, and Reproduction. Specifically, the existing practices could not guarantee Requirement B.1.b:

The preserver should be able to demonstrate that the procedures and system(s) used to transfer records to the archival institution or program; maintain them; and reproduce them embody adequate and effective controls to guarantee the records’ identity and integrity, and specifically that security and control procedures are implemented and monitored…

---


Rather than RLG’s *An Audit Checklist for the Certification of Trusted Digital Repositories*, Schmidt used the *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist* to perform an audit on the existing preservation practices of the H-Net e-mail list archive. Published in February 2007 by the Center for Research Libraries (CRL) and OCLC, the TRAC is an objective, prescriptive methodology to establish the trustworthiness of a digital repository based on the earlier draft document. The checklist consists of 84 audit criteria organized into three sections: Organizational Infrastructure; Digital Object Management; and Technologies, Technical Infrastructure, and Security. Each section is divided into subsections of audit criteria to compare to current local capabilities. This comparison, known as a “gap analysis,” is the difference between the current state and the desired state of a criterion governing the trustworthiness of a repository. Once the gap has been identified, strategies may be formulated that will narrow it.

The audit revealed that the H-Net archive was lacking in many best practices in preservation as described above. As the instructions for using the TRAC checklist note the preference for written documentation as evidence of support for the criteria, the lack of comprehensive written preservation policies marked the H-Net archive as a deficient digital repository. Conducted in February 2008 and published on the project website in March 2008, the checklist containing the original TRAC assessment may be found at [http://www.h-net.org/archive/documentation/TRAC%20current%20publish.pdf](http://www.h-net.org/archive/documentation/TRAC%20current%20publish.pdf).

### 2.3. Preservation Improvement Plan

2.3.1. Fixity measures
As both the individual messages and notebook files are archival information packages as defined by OAIS—AIPs and AICs comprised of AIPs, respectively—both require the means to establish and ensure their integrity and authenticity. The improvement plan thus suggested that the MD5 hashes created for message discovery purposes be used to perform checksums to establish and validate fixity for individual messages. Rather than waiting until a notebook file is complete, MD5 hashes would be assigned within 24 hours of a message posting and stored in a database. (Unfortunately, the current version of the LISTSERV software would not allow the assignment of hashes at the actual time of posting. Boone wrote a letter to L-Soft suggesting that the software be opened to this type of customization in a future release.) At the end of the seven-day period when a notebook closes, the MD5 hashes would be validated.

Notebook files would be assigned SHA-2 hashes on creation. On the recommendation of the National Institute of Standards and Technology (NIST), MATRIX has adopted the SHA-2 message digest algorithm for use in establishing and verifying fixity rather than the more commonly used MD5 algorithm. (Data collisions have occurred with MD5, indicating that it may not guarantee file integrity.) Notebooks in existence at the time these measures were implemented would also be assigned SHA-2 hashes. Validation of the SHA-2 hashes would occur weekly.

Any errors encountered during validation of messages or notebook files would be logged, investigated, and corrected manually as appropriate.

2.3.2. Accurate message creation metadata
From an archival standpoint, the loss of the metadata for the original author and creation date when a message is edited destroys some of that message’s provenance. Rather than relying on editors to add back the metadata, Boone suggested the development of a new web-based list editing interface that would keep the original metadata along with new metadata from later actions performed on the message.

2.3.3. Administrative restrictions
To eliminate the loophole that would allow list editors to change or delete notebooks from the H-Net file system, Boone suggested removing the editors’ ability to retrieve and change notebook files. Strategies to overcome the possible security threats to H-Net from MATRIX employees
holding root accounts were entertained, including limiting access permissions and/or developing the means to generate audit logs of non-standard activity.

2.3.4. Preservation of attachments on private lists
At the time of the original assessment, only a subset of the private lists were browsable in the manner of the public lists. In order to make preservation of the private lists worthwhile, access would need to be provided to them through constructed URLs in the manner of the public lists, with a download link provided in the browse list.

In order to determine which formats might require migration over time, an inventory of attachments needed to be taken. A technology watch would 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 necessary, file format conversion tools for the most popular formats could be kept in reserve on the H-Net website, or links provided to other websites that offer such tools. Alternatively, MATRIX could provide the means for automatic conversion to the new file formats.

2.3.5. Preservation of links to content
An advisory board member had suggested preserving links to content within messages, to protect the research value of the H-Net messages against a future of broken or dead links. Boone agreed to investigate redirecting links to their original websites archived in the Wayback Machine of the Internet Archive (www.archive.org). As the Wayback Machine is not infallible, it was understood that links to every bit of original content might not be possible.

2.3.6. Shorter persistent URLs
From a better usability standpoint, another advisory board member suggested shortening the long persistent URLs for messages. Shorter persistent URLs would be more inviting for use in citations—particularly in printed materials—and less prone to error. To that end, Schmidt recommended that Boone look into creating and mapping the long persistent message URLs to shorter, more user-friendly alternatives.

2.3.7. Backup and archiving
Improvements to the MATRIX backup system and a plan to create archival copies of H-Net were articulated in the supplemental report MATRIX and H-Net Backup and Archival Storage,
In addition to MATRIX’s current backup system, two offsite backup plans were suggested. One was to set up a reciprocal storage arrangement with the Inter-University Consortium of Political and Social Research (ICPSR) at the University of Michigan in Ann Arbor. This mirroring arrangement would entail daily synchronization and backup of MATRIX data to ICPSR servers, and MATRIX would provide the same service for 3 TB of ICPSR’s data. The second offsite storage recommendation involved moving the “permanent” backup tapes to a commercial storage provider in Lansing. The Michigan State University Archives & Historical Collections offered to provide this service. Also, Schmidt recommended that the “permanent” backup tapes be assigned the designation of “long-term” and put on a retention schedule.

Schmidt also recommended the creation of archival copies of H-Net, separate from MATRIX backup data. These copies would be created on an annual basis. For the near future, it was recommended that the copies be made to LTO tape. Copying the archive over to a server-based repository would be considered for the future, as would participation in LOCKSS or another distributed preservation arrangement.

2.3.8. Creation of digital preservation policies
During the last quarter of 2008, Schmidt decided to model digital preservation policies for the H-Net e-mail lists on a digital preservation management methodology developed by Nancy McGovern and others at Cornell University and ICPSR. This methodology provided templates for the development of a policy framework, processes, and procedures for a digital repository based on the OAIS model and the requirements outlined in the RLG-OCLC Trusted Digital Repositories document. The creation of digital preservation policies would fulfill the TRAC requirement of documenting the preservation activities of the H-Net list archive.

2.4. Archival Advisory Board, H-Net Council, and H-Net Editorial Board Review
The archival advisory board for the H-Net e-mail list preservation project included: Charles Dollar, professor emeritus of the School of Library, Archival, and Information Studies at the University of British Columbia, Vancouver; Patricia K. Galloway, associate professor of the School of Information, University of Texas at Austin; Margaret Hedstrom, associate professor in the School of Information and faculty coordinator of the Archives and Records Management
specialization at the University of Michigan, Ann Arbor; Michael Nelson, assistant professor of computer science at Old Dominion University; and Caryn Wojcik, government records archivist for the State of Michigan. Schmidt sent all of the documentation created for the H-Net e-mail list preservation project to this advisory board for review. Galloway in particular provided useful feedback during the course of the project, and Dollar, Nelson, and Wojcik also offered helpful advice.

Likewise, Dr. Peter Knupfer, director of H-Net, was regularly apprised of activity on the project. Dr. Knupfer brought the suggested technical improvements to the H-Net Council and editorial board for review and approval during a Fall 2008 meeting.

2.5. Implementation of the H-Net Preservation Plan
After the meeting of the H-Net Council and editorial board, Schmidt and Boone moved forward in implementing the following technology and policy documentation improvements. Most of the suggested improvements were approved, and implementation took place over the course of 2009.

2.5.1. Fixity
Boone augmented the H-Net e-mail list archive with the means to establish and validate fixity for both individual messages and notebook files, as had been suggested. Instead of using the extant MD5 hashes as checksums for messages, however, SHA-2 hashes were used to better ensure authenticity and integrity. The MD5 hashes are still being used for message discovery purposes. Figure 5 shows the H-Net information packages and how they relate to the new fixity measures.

2.5.2. Web-based list editing interface for accurate message provenance
The H-Net Council decided against the development of a web-based list editing interface to ensure accurate message provenance. Legacy messages would not benefit from this interface, and the Council wished to focus the scant available development resources on making other improvements to H-Net, such as adding social networking functionality.

2.5.3. Administrative restrictions
Boone was able to eliminate list editors’ ability to retrieve and change notebook files. A risk analysis of the consequences of limiting the H-Net access restrictions of MATRIX staff, however, resulted in the decision to make no changes. Several staff members must hold root account privileges to ensure 24/7 availability of all of the systems hosted by MATRIX, such as online
history courses to which students must have ongoing access for course fulfillment. Furthermore, restricting root account access to LISTSERV administration would be difficult if not impossible, and developing the means to generate audit logs of non-standard activity would not be a worthwhile use of resources. Although H-Net is a valuable scholarly resource, the nature of the information 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, and the risk of them deliberately compromising the H-Net lists is minimal to nonexistent.

**Figure 5. H-Net Information Packages**

![H-Net Information Packages Diagram](image)

**2.5.4. Preservation of attachments**

Boone demonstrated that the H-Net private e-mail lists may be made browsable with links to attachments. As the lists are private, subscribers who are unfamiliar with LISTSERV commands need special instructions to browse a list archive. H-Net administration has been advised to provide current private list subscribers with this information and to rewrite list welcome messages to include that information for new subscribers.

An inventory revealed that attachments are found in less than 0.01 percent of all H-Net messages. MATRIX thus decided against implementing a format migration or normalization plan at this
Most attachments are in Office, PDF, JPEG, and other common formats, which should open in viewers, later versions of the applications, or other applications. MATRIX will assist any users who report problems opening attachments.

2.5.5. Preservation of links to content

Although Boone explored the idea of redirecting links within messages to original web pages as archived in the Wayback Machine, MATRIX did not move forward with it at this time.

2.5.6. Creation of shorter persistent URLs

MATRIX decided against creating shorter persistent URLs for messages using Shrinkster or Tiny URL redirects. As this added a layer of complexity and was not of direct preservation value, it was ultimately not considered a worthwhile use of development resources.

2.5.7. Offsite backup for MATRIX and archival storage solution for H-Net

In terms of the recommended backup improvements, MATRIX entered into the reciprocal storage arrangement with ICPSR. Instead of using an off-campus storage facility, the MSU Archives has agreed to store MATRIX’s long-term backup tapes. These tapes are now on a two-year retention schedule.

MATRIX will create and maintain archival copies of the H-Net e-mail list data separate from its other files and applications. On an annual basis, MATRIX will make two copies of H-Net messages posted during the previous calendar year, along with associated metadata and supporting documentation, onto archival-quality LTO tapes using GNU Tar archiving software. The first archival copy created contains all messages posted from 1989-2008; subsequent archival copies will be created annually. One tape will be stored at the MSU Archives and the second tape kept in a secure location on the MATRIX premises. Media refreshment for the tapes is scheduled for every five years.

MATRIX recognizes that tape is not an ideal digital archival storage medium, and this solution precludes the practicality of running regular file integrity checks. Thus, MATRIX will explore storing the archival copies of the H-Net lists in a server-based digital repository. As part of a future greater Michigan State University data management and preservation initiative, MATRIX and H-Net may also eventually participate in a distributed archival storage system or enter into an inter-institutional LOCKSS- or MetaArchive-like distributed archiving arrangement.
2.5.8. Documentation of digital preservation policies

Using the Cornell/ICPSR digital preservation management framework template as a guideline, Schmidt documented digital preservation policies for the H-Net e-mail lists. These policies were published internally at MATRIX and on the H-Net preservation project website in July 2009, and may be accessed at http://www.h-net.org/archive/doc.php under “H-Net Digital Preservation Policies and Procedures.”

2.6. Longer-Term Needs and Migration Strategy

Implementation of the preservation strategies noted above will help to ensure the longevity of the content of the H-Net e-mail archive. For the longer term, MATRIX must remain vigilant with ongoing TRAC evaluations and digital preservation policy reviews, exploring participation in university-wide and regional preservation systems and repositories, and complete the process of identifying a successor host for the H-Net lists. Note that file format and media migration strategies for the H-Net e-mail list archive were discussed in previous sections. The “migration strategy” mentioned here is assumed to refer to the need for a succession plan for the H-Net list archive.

2.6.1. Ongoing TRAC evaluations

A second TRAC assessment was performed on completion of the policies in July 2009 and posted to the project website at http://www.h-net.org/archive/trac7-09.pdf. Most of the gaps uncovered in the earlier evaluation had indeed narrowed, bringing the H-Net e-mail list archive closer to achieving the status of a trusted digital repository for the valued academic discourse contained in its discussion networks. In fact, TRAC criterion A3.9 states that a trusted digital repository must commit to regular audits, and the new digital preservation policy framework for the H-Net list archive calls for audits to be conducted every two years.

2.6.2. Periodic digital preservation policy reviews

As with the TRAC evaluations, digital preservation policy reviews must be conducted periodically and the policy framework and supporting documentation updated accordingly. The digital preservation landscape is undergoing rapid change, and new strategies, technologies, and financial considerations will necessitate changes in policy over time. Per the current policy framework, reviews should take place every two years.
2.6.3. Participation in university-wide and regional distributed preservation systems

MATRIX should strive to participate in a distributed storage system that offers preservation functionality. Within the next two-to-five years, Michigan State University plans to implement the Integrated Rule-Oriented Data System (iRODS) or a similar rules-based preservation system that MATRIX could join. Schmidt is currently spearheading a digital curation planning project for the university, which should help create a foundation for establishing an institutional repository or a federation of digital repositories on the campus. The Committee on Institutional Cooperation (CIC), which includes Michigan State University and ten other universities in the Upper Midwest, has plans to implement a regional repository which MATRIX could participate in. Another option might be participation in a LOCKSS- or MetaArchive-based system, as articulated in an earlier section.

2.6.4. Succession plan

Before receiving the NHPRC grant, MATRIX had made preliminary inquiries with the Library of Congress and OCLC as alternative long-term repositories for the H-Net e-mail list archive. No further movement forward on these negotiations for a succession plan has occurred, however. TRAC criterion A1.2, states: “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.” To close this preservation gap, MATRIX must identify, negotiate with, and make preliminary plans with a potential successor. MATRIX can narrow the gap in the meantime by articulating the qualities required in a successor host and documenting the intent to put a succession plan in place.
3. Enhancing Access to H-Net Through Improved Search Tools

The research into Semantic Augmented Consensus Clustering (SACC) techniques was to improve access to H-Net messages by enabling thematic search capabilities. Through examination of the semantic relationships between words in an H-Net message (document), SACC techniques could identify semantically close words that represent the theme of that message. The research would then test whether fewer pertinent words could be used to represent the message for search or clustering, providing more efficient search processes and more useful results. Unfortunately, this SACC research met with mixed, limited success.

A dedicated computer science graduate student tested multiple variations on ways to determine and utilize sense in the clustering process. Using an algorithm that incorporated the Wordnet ontology, close nouns in a document were identified and documents clustered based on the distance between the nouns, a technique known as “sense clustering.” Dr. Punch planned to use either the “sense clustering” technique alone or in conjunction with other approaches via ensemble clustering, showing how much “sense” matters in improving the quality of the clusters discovered.

Variations on the Wordnet algorithm were run using two different sets of data: Reuters data and the “newsgroup 20” data set (http://people.csail.mit.edu/jrennie/20Newsgroups/). These data sets pre-clustered the documents, providing a ground truth for comparison to the standard Latent Semantic Indexing (LSI) and k-means approaches to document clustering. While some improvement in clustering documents from the newsgroup 20 data set was seen when sense information was included, the same techniques yielded little improvement in the Reuters data. Anomalies observed in the intra-document clustering did indicate opportunities to improve the algorithm.

Research activity on this aspect of the H-Net preservation/access project was suspended after the first quarter of 2009. The techniques had not been tested on H-Net e-mail records as originally planned, as the prototypes were not yet efficient enough to scale up to 1,000,000 records. (Note that the testbed comprised 10,000 records.) If more grant funding may be obtained from the National Science Foundation (NSF) or another source, research may continue on two major fronts: Better ways to incorporate sense in clustering, and the development of tools to better measure the information gain/change from co-clustering methods.
4. Project Information Dissemination

During the course of the H-Net e-mail list preservation and access project, information about its progress was disseminated through a project website, published articles, and presentations at conferences and to groups of graduate students. These activities fulfilled the requirement to “seek interaction with a border interested community of colleagues in computer science, scholarly networking and discussion lists and forums, digital library, and electronic record and archivist communities,” as described on p. 16 of the original grant proposal.

4.1 Project Website

A website for this project, http://www.h-net.org/archive, went live during the first half of 2007. Project documentation, progress reports, presentations, and papers, as well as links to policies and other relevant content, were posted to the site as they were created.

4.2 Articles and Presentations

During the course of the project, articles were developed for publications online and in key journals, and presentations were delivered at strategic conferences and meetings.

4.2.1 Articles

- “Preserving the H-Net E-Mail Lists: A Case Study in Trusted Digital Repository Assessment” (working title) by Lisa M. Schmidt, American Archivist, submitted December 2009, pending publication acceptance by March 2010 (will submit to Archival Issues, the journal of the Midwest Archives Conference, if not accepted by AA)

4.2.2 Presentations

- “Preserving a Born-Digital Archive: The H-Net E-Mail Lists” presented by Lisa M. Schmidt at the University of Texas School of Information, Austin, TX, November 2009
• “A Sustainable Preservation Plan: The H-Net E-Mail Lists” presented by Lisa M. Schmidt at the Society of American Archivists (SAA) Annual Meeting, Austin, TX, August 2009

• “Electronic Mailing List Preservation Takes Off: The H-Net Archive” presented by Lisa M. Schmidt at the Midwest Archives Conference (MAC), St. Louis, MO, May 2009

• “Preserving Electronic Mailing Lists: The H-Net Archive” poster presented by Lisa M. Schmidt at the Digital Curation Curriculum (DigCCurr) Symposium, Chapel Hill, NC, April 2009

• “Digital Preservation Concepts in Action: Assessing the H-Net Archive as a Trusted Digital Repository” presented by Lisa M. Schmidt to the “Practical Engagement Workshop in Digital Preservation” class at the University of Michigan School of Information, Ann Arbor, MI, March 2009


• “Preserving the H-Net E-Mail Lists: A Case Study in Trusted Digital Repository Assessment” presented by Lisa M. Schmidt to the digital preservation interest group at the American Library Association (ALA) Midwinter Meeting, Denver, CO, January 2009

• H-Net preservation project presentation presented by Lisa M. Schmidt at the Committee on Institutional Cooperation University Archivists Group (CIC-UAG) Forum, Michigan State University, East Lansing, MI, September 2008


• “Taming the Wild LISTSERV; or, How to Preserve Specialized E-Mail Lists” presented by Lisa M. Schmidt at the Society of Southwest Archivists (SSA) Annual Meeting, Houston, TX, May 2008

• “H-Net and Scholarly Discourse in the Digital Age: A New Approach to Data Mining Email Discussion Lists” presented by Mark Kornbluh, William Punch, and Wayne Dyksen at the Chicago Colloquium on Digital Humanities and Computer Science, Chicago, IL, November 2006
5. Conclusions

The NHPRC-funded H-Net e-mail list archive project succeeding in meeting the goal of ensuring authenticity, preservation, and persistence of messages, and the results should prove useful to those managing large collections of electronic records. Unfortunately, the research in improving access through clustered search techniques fell short of producing useful results.

As the first scholarly exploration of e-mail list preservation, this study is directly applicable to the preservation of other e-mail lists, particularly those based on LISTSERV. Also, the tools used in this study—such as the TRAC and the Cornell/ICPSR digital preservation management methodology—can be applied to more complex data sets than the relatively small, mostly homogenous H-Net lists. Finally, Schmidt allowed researchers on the CERP e-mail preservation study to test their XML parser tool on H-Net list data, providing an intriguing angle on future preservation and access possibilities.

5.1. Applicability to E-Mail List Preservation Projects

The success of the H-Net list archive preservation project may be extended to the analysis and preservation of other e-mail list archives, particularly those base on LISTSERV. Not all e-mail lists contain information considered of long-term research value, of course. Many lists seem to focus on ephemeral topics relevant only to the moment a message posts. The argument may be made, however, that such lists document social and cultural trends. For example, even though the Society of American Archivists’ Archives & Archivists (A&A) discussion lists do not document the official business of SAA, that they recorded evidence of matters of concern to subscribers at discrete points in time might deem them preservation-worthy. At any rate, the H-Net list preservation project offers other list providers a blueprint for preserving their own e-mail list archives.

5.2. Applicability to Complex Digital Preservation Projects at Michigan State

Other digital preservation projects at Michigan State University involving proprietary file formats and complex digital objects will leverage the knowledge gained through the H-Net list archive preservation project. Schmidt plans to document digital preservation policies for MATRIX using the Cornell/ICPSR digital preservation management policy framework. On a larger scale, she is engaged in a university-wide digital curation planning project expected to result in best practices guidelines and to provide the foundation for the development of an institutional repository. Also, the MSU Archives recently received NHPRC funding to develop the Spartan Archive, an OAIS-
based electronic records repository that is expected to leverage the expertise behind the H-Net list preservation project.

5.3. Possibilities for Improved Search Techniques

During the course of this project, CERP researchers tested their XML parser, used to archive e-mail created in proprietary formats, on H-Net list data. Riccardo Ferrante, electronic records program director at the Smithsonian Institution Archives and CERP principal investigator, reported success in preserving the H-Net weekly notebook files using the XML parser and the CERP e-mail account schema. As noted previously, the plain text file format of H-Net precludes the need for conversion to XML, as the data is already in an archival format. XML affords more benefits than its status as an archival format, however, as it provides structure to data that can make it more easily searchable. To date, a search framework that leverages the information encoded in the CERP e-mail account schema does not exist. Once such a framework is developed, conversion of H-Net e-mail data to the XML format may become a more attractive preservation option.
6. Reference List


