

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

**Interim Narrative Progress Report, February 1, 2008–July 31, 2008**

**Project Activities Undertaken**

Of the project activities scheduled, MATRIX undertook the following during the reporting period:

- Document existing authenticity, preservation, and persistence practices of H-Net discussion network records, including evaluation of existing preservation practices using the NARA/OCLC Trustworthy Repositories Audit & Certification: Criteria Checklist (TRAC)
- Post document describing existing practices on participants' section of website, including the TRAC
- Evaluate with Archival Advisory Board the progress on workplan for H-Net preservation and make any recommended revisions
- Deliver presentation reporting on preliminary project results at Society of Southwest Archivists (SSA) annual meeting
- Draft workplan for implementing H-Net preservation plan during next 12 months, to include revisions based on the TRAC
- Semantic Augmented Consensus Clustering (SACC) research on H-Net records



COLLEGE OF  
SOCIAL SCIENCE  
MATRIX

Michigan State University  
310 Auditorium Building  
East Lansing, Michigan  
48824-1120

517/355-9300  
FAX: 517/355-8363  
<http://www.matrix.msu.edu>

***Document existing authenticity, preservation, and persistence practices of H-Net discussion network records, including evaluation of existing preservation practices using the NARA/OCLC Trustworthy Repositories Audit & Certification: Criteria Checklist (TRAC).*** During the reporting period, electronic records archivist Lisa Schmidt worked with MATRIX systems administrator Dennis Boone to finish documenting existing H-Net processes and preservation practices. This included a thorough evaluation of existing preservation practices using the TRAC, covering the areas of organizational infrastructure; digital object management; and technologies, technical infrastructure, and security.

***Post document describing exiting practices on participants' section of website, including the TRAC.*** In March 2008, the documents "Preservation of the H-Net E-Mail Lists: Current Practices" and the "Trustworthy Repositories Audit & Certification: Criteria Checklist for H-Net" were completed and posted to the "Documentation" page of the project website. (<http://www.h-net.org/archive/doc.php>).

***Evaluate with Archival Advisory Board the progress on workplan for H-Net preservation and make any recommended revisions.*** In March 2008, Ms. Schmidt sent the documents "Preservation of the H-Net E-Mail Lists: Current

Practices” and the “Trustworthy Repositories Audit & Certification: Criteria Checklist for H-Net” to the Archival Advisory Board for review and suggestions for improvement. Board members noted that the current system functions very well as a preservation system, though improvements are needed to make it more archivally sound. Major areas of improvement to be considered include:

- Backup and storage—including provision for another set of backup tapes or a server mirror, more secure storage systems, a backup log, and use of a distributed storage system
- Authenticity—including shortening and standardizing the ingest time window, defining and documenting access permissions, maintaining an audit log of record activity, and performing regular fixity checks.
- Generation of a shorter persistent URL for use in bookmarking and citations
- Maintaining the accuracy of creation metadata when an editor makes changes to a message
- Migration plan for attachments on private H-Net lists, which includes making those lists browsable
- Compliance with TRAC criteria—including (but not limited to) establishing a succession plan, establishing or leveraging a technology watch, and fully documenting such aspects of the preservation plan as technology history; change management; staff roles, responsibilities, and authorizations; and a disaster recovery strategy

***Deliver presentation reporting on preliminary project results at Society of Southwest Archivists (SSA) annual meeting.*** On May 23, 2008, Ms. Schmidt delivered the presentation “Taming the Wild LISTSERV; or, How to Preserve Specialized E-Mail Lists” at the SSA conference in Houston. The presentation covered the project to date, including an overview of current preservation practices and suggested improvements. This presentation was part of a session entitled “Archiving Born-Digital Records and Manuscripts in University Settings” moderated by Archival Advisory Board member Pat Galloway. The session generated much interest, with more than 40 people in attendance. Ms. Schmidt’s presentation has been posted to the H-Net archive project website (<http://www.h-net.org/archive/presentations.php>).

***Draft workplan for implementing H-Net preservation plan during next 12 months, to include revisions based on the TRAC.*** Ms. Schmidt and Mr. Boone are in the process of putting together an H-Net preservation improvements workplan. A list of recommended requirements for improving H-Net preservation practices will include action items and target dates for completion. This list and an explanatory document will be circulated for approval to the Archivist Advisory Board, MATRIX and H-Net management, and the H-Net Council early in the next reporting cycle.

***Semantic Augmented Consensus Clustering (SACC) research on H-Net records.*** Bill Punch, Associated Faculty at MATRIX and Associate Professor of Computer Science and Engineering at MSU, reports that results of the continued SACC research have been mixed.

A dedicated computer science graduate student has been exploring and evaluating different approaches to the problem, building and testing prototypes to determine the best clustering methodology. Using an algorithm that incorporates the Wordnet ontology, close nouns in a document are identified and documents clustered based on the distance between the nouns, a technique known as “sense clustering.”

Variations on this algorithm have been 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 have the documents pre-clustered, providing a ground truth for comparison to the standard Latent Semantic Indexing (LSI) approach to document clustering. Dr. Punch plans 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.

While some improvement in clustering documents from the newsgroup 20 data set has been seen when sense information is included, the same techniques have yielded little improvement in the Reuters data. Anomalies observed in the intra-document clustering indicate that there are opportunities to improve the algorithm.

### **Additional Project Developments**

In addition to activities scheduled on the project workplan, Ms Schmidt managed the following developments during the reporting period:

- Participation in a National Information Standards Organization (NISO) Digital Preservation Forum in DC in March 2008, the Midwest Archives Conference (MAC) annual meeting in Louisville in April, and the Michigan Archival Association (MAA) annual meeting in Mackinaw City, MI in June. All afforded educational and collaboration opportunities that can be leveraged for the success of the H-Net archives preservation project, especially the MAC meeting.
- Submission and acceptance of a proposal to present on the project at the Research Forum of the Society of American Archivists (SAA) conference in August 2008.
- Invitation to speak to the Digital Preservation section of the American Library Association (ALA) during that organization’s next mid-winter meeting. The section chair is particularly interested in MATRIX’s work with the TRAC.
- Participation in a week-long, SAA-sponsored “Electronic Records Summer Camp” workshop at the San Diego Supercomputer Center (SDSC) at the University of California, San Diego. The workshop focused on a hands-on introduction to the integrated Rule-based Data System (iRODS) data management cyberinfrastructure and its applicability to digital preservation. Rules written in iRODS could operationalize archival preservation policies, such as those suggested by the TRAC criteria. The iRODS team has asked to test the system with the H-Net archive. This iRODS technology may eventually prove to be an important plank in the H-Net preservation strategy as well as the preservation strategies for MATRIX and the Michigan State University Archives.

- Request from the Smithsonian Institution, a partner with the Rockefeller Archive Center in the Collaborative Electronic Records Project (CERP), to test their e-mail preservation technology on H-Net files. This cooperative effort may yield useful results for the H-Net archives preservation project.



Mark Kornbluh  
Principal Investigator



Lisa M. Schmidt  
Project Manager