Taming the Wild LISTSERV; or, How to Preserve Specialized E-Mail Lists

By Lisa M. Schmidt • MATRIX: The Center for Humane Arts, Letters, and Social Sciences Online, Michigan State University

H-Net: Humanities and Social Sciences Online, an international consortium of scholars and teachers, is the oldest collection of born-digital, contemporaneously generated and content-moderated arts, humanities, and social science material on the Internet. A valuable scholarly resource, H-Net includes more than one million e-mail messages on 180 public and 230 private lists. H-Net is hosted by MATRIX: The Center for Humane Arts, Letters, and Social Sciences Online, a digital humanities research center at Michigan State University.

MATRIX received a grant from the National Historical Publications and Records Commission (NHPRC) to conduct an assessment of existing preservation policies and practices for H-Net and to develop an improved long-term preservation plan. This includes applying the NARA/OCLC Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) to H-Net. The work on H-Net preservation will be useful to archivists and others who manage large collections of electronic records.

H-Net uses five out of thirteen MATRIX servers, with one server providing storage and backup for the network configuration. Incremental tape backups are performed daily, a full backup weekly. Tapes are stored in a secure location and replaced as needed, usually when a cartridge breaks. A full permanent backup is performed monthly, with those tapes kept in a minimally secured room.

H-Net runs on LISTSERV software, which distributes messages to editors and subscribers and allows administrators to create and add lists. All messages must be written in plain text, and no attachments are allowed on the public lists. A subscriber sends a message to an editor who either approves or edits it before posting. In the latter case, it essentially becomes a new message; if the editor does not manually add back the original author's name and message creation date, that metadata is lost. The posting process can take from a few seconds to several days. Once a message posts, it becomes part of a flat text file, or "notebook." A notebook includes messages posted during a weekly time period.

This is a condensed version of the paper delivered as part of Session 10: Archiving Born-Digital Records and Manuscripts in University Settings.

Every 24 hours, the newest messages in the current notebook file are copied to a proprietary BRS database, where they are available for full-text search. As a separate operation, a log browse cache application reads the notebook messages and creates MD5 hashes for each message seven days after the last message posts to a notebook. A cache builder script then writes message metadata to a MySQL database cache. This includes the filename of the notebook where the message is stored; the offset, or byte position in the notebook file; name and e-mail address of the "author"; the subject line; the date in two formats; and the MD5 hash, or "messageid."

When a user browses an H-Net list and selects a message, a log browse application pulls it from the original notebook file and builds a URL that combines its filename and MD5 hash. This URL serves as a persistent identifier that can be bookmarked for reference and citation purposes.

The MATRIX backup and storage processes provide one piece of the current strategy for preserving the H-Net lists. The most significant property of the messages that must be preserved is their content, and most of them are written in plain text ASCII and UTF-8—recommended nonproprietary, archival formats for text. Authenticity is based on the author and/or editor of a message informally checking it after posting. Also, if a user receives a broken URL when attempting retrieval, the authenticity of the message has been compromised. The cached metadata fulfills the requirements for preservation description information (PDI) for each information package, as recommended by the Open Archival Information System (OAIS) model.

For better backup and storage, MATRIX must implement a regular media refreshment schedule for all tapes. More than one set of permanent backup tapes should be made and stored offsite, or a server mirror should be established. Storage of the tapes must be more secure, and a backup log must be maintained. In addition, MATRIX should participate in a distributed storage system such as LOCKSS or the San Diego Supercomputer Center's iRODS.

H-Net is missing the authenticity boat. The time window from when an editor approves a message to when it posts must be shortened to seconds rather than weeks. Access permissions must be defined and documented. Audit logs should track activities associated with records. If an editor makes any changes to an original message, their metadata should be automatically added to that of the author.

Regular integrity checks should be performed, with a message digest assigned at ingest, new messages verified weekly, and regular fixity checks performed quarterly. Consideration should be given to using the SHA-2 message digest algorithm, as the integrity of MD5 has been compromised. MD5 could still be used to calculate the name of the message. The current persistent URL is too long, however; it should be mapped to a shorter URL for use in citations.

No migration strategy is required for the messages and notebooks, as they are in stable, open, plain text formats. The attachments on the private lists must be detached and stored separately, with conversion to current formats provided on demand.

Applying the TRAC checklist to the current H-Net preservation practices and policies revealed a number of other measures that must be taken to ensure a more archivally sound system. These include:

- Developing a succession plan, in the event that MATRIX can no longer host H-Net
- Determining a periodic review or trigger event definition to ensure responsiveness to technological developments and evolving requirements
- Establishing a technology watch
- Documenting H-Net's technology history, a change management system, staff roles and authorizations, and a written recovery plan.

For more on this project, see http://www.h-net.org/archive/.

Southwestern Archivist

Needs You!

Tell your colleagues about your acquisitions, projects, exhibits, or grants — submit your repository news by October 10th.

Photographs (300dpi in a native image format) are highly encouraged. Be sure to provide the caption / credit informaton that you want to accompany the image!