

An In-Depth Look at Information Packages in the H-Net Preservation System

**Preservation of the H-Net E-Mail Lists
Supplemental Report**

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

An In-Depth Look at Information Packages in the H-Net Preservation System

H-Net: Humanities and Social Sciences Online is an international consortium of scholars and teachers that has grown to include more than 180 scholarly social sciences and humanities networks hosted by MATRIX: Center for Humane Arts, Letters and Social Sciences Online at Michigan State University. In addition to its public lists, H-Net includes more than 230 “private” lists used by editors, council members, and administrators for planning, testing, and advisory purposes. MATRIX received a grant from the National Historical Publications and Records Commission (NHPRC) to advance the state of e-mail preservation by assessing and improving upon the digital preservation practices for the H-Net electronic mailing lists to ensure longevity of the content. Representing a compilation of years of academic discourse, with messages bookmarked and cited in scholarly research and publications, the H-Net lists are considered a valuable scholarly resource requiring long-term preservation.

The document “ Preservation of the H-Net E-Mail Lists: Current Practices,” posted on the H-Net archive project website in spring 2008, provides a description and analysis of the H-Net preservation system hosted by MATRIX as well as some suggestions for improvement.¹ (A more comprehensive document covering suggestions for improvements was posted in August 2008.²) The analysis includes a high-level

¹ Schmidt, Lisa M., “Preservation of the H-Net E-Mail Lists: Current Practices,” H-Net: Preserving and Improving Access to Specialized Electronic Mailing List Archives, March 2008, <http://www.h-net.org/archive/documentation/H-Net%20Current%20Practices%20Post2.pdf>.

² Schmidt, Lisa M., “Preservation of the H-Net E-Mail Lists: Suggested Improvements,” H-Net: Preserving and Improving Access to Specialized Electronic Mailing List Archives, August 2008, <http://www.h-net.org/archive/documentation/hnetpresimprov.pdf>.

mapping of the H-Net message ingest, storage, and retrieval processes to the Open Archival Information System (OAIS) model.³

This supplemental report provides a more in-depth analysis of the mapping, particularly as regards the components of the Information Packages (IPs)⁴ and assumes the implementation of suggested improvements regarding fixity; these improvements are scheduled for completion no later than the first quarter of 2009. In the course of conducting this analysis, some IPs have been reclassified from their identifications in earlier documents and presentations posted on the H-Net archive project website. The reclassifications supersede the earlier identifications.

Note that this report assumes familiarity with the H-Net system. For a complete description of the message posting (ingest), storage, and retrieval processes of H-Net, please refer to the “Current Practices” document.

H-Net Public List Information Packages

H-Net public list messages and their accompanying metadata are present in the system as the three standard Information Package variants: Submission Information Packages (SIPs), Archival Information Packages (AIPs), and Dissemination Information Packages (DIPs). Seven-day concatenations of messages gathered into “notebook” files and accompanying preservation metadata are AIP specializations known as Archival Information Collections (AICs). (See Figure 1.)

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

⁴ CCSDS, 2-6 – 2-7.

Submission Information Packages (SIPs)

Submission Information Packages (SIPs) in the H-Net preservation system consist of the messages posted by the editors. LISTSERV has stripped out the header information peculiar to the sender's mailer software and used for mailing purposes. Messages include the body of the message and the following header fields:

- Date—date and time message sent
- Reply-To—list name and e-mail
- Sender—list name and e-mail
- From—name and e-mail of individual sender
- Subject
- Mime-Version—such as “1.0 (Apple Message framework v752.3)”
- Content-Transfer-Encoding—such as “7bit”
- Content-Type—such as “text/plain; charset=US-ASCII; delp=yes; format=flowed”

Archival Information Packages (AIPs)

Within 24 hours of submission, metadata is extracted and MD5 hashes created for each message. This metadata, which is stored in a log browse cache database and used for more expeditious message retrieval, includes:

- filename—name of notebook file where message is stored
- offset—byte position in notebook file where message is stored
- from—name and e-mail address of original author or default to editor's information
- subject—as posted by the editor
- dpb—date posted
- cbd—date in a different format for sorting purposes
- messageid—MD5 hash

The MD5 hashes are also stored in a separate database used to run fixity checks on the messages.

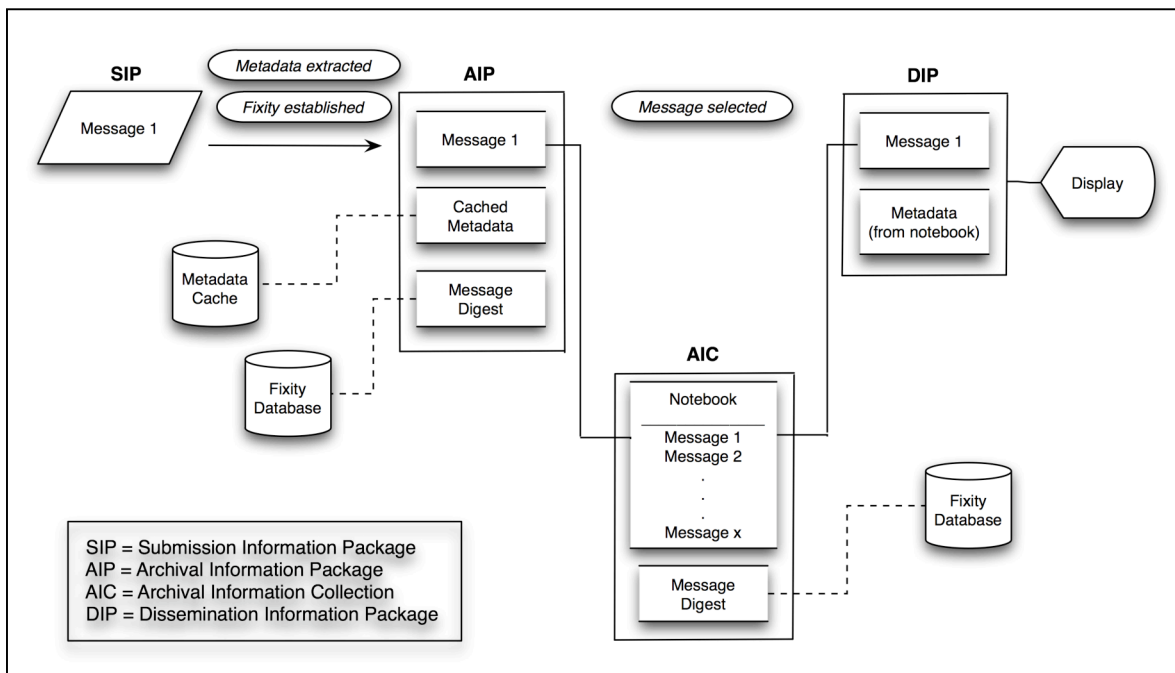
Archival Information Packages (AIPs) thus consist of the message ingested (body and header), the metadata for that message stored in the cache, and the MD5 hash for the message stored in the database of fixity information.

The OAIS model also describes AIP specializations, including a collection of AIPs known as an Archival Information Collection (AIC).⁵ In the H-Net system, notebook files consisting of a seven-day accumulation of messages for a given list make up part of an AIC. The AIC also includes a SHA-2 hash assigned to the notebook and stored in a database of fixity information for notebook files.

Dissemination Information Packages (DIPs)

When a user selects a message through the browser interface, the system retrieves it through reference to the metadata stored in the cache. The selected message and metadata from the notebook make up the Dissemination Information Package (DIP).

Figure 1. H-Net Information Packages



⁵ CCSDS, 4-37 – 4-39, 4-42 – 4-43.

H-Net Private List Information Packages

Some messages on the private, administrative H-Net lists contain attachments and could be handled differently in the preservation process. A copy of the attachment could be generated at the time of ingest and normalized into an open source or de facto standard format for greater accessibility. The original message (including its embedded attachment), related cached metadata, and its fixity information would be considered a specialized AIP known as an Archival Information Unit (AIU),⁶ as would the normalized attachment in combination with its own fixity metadata. The AIUs plus context metadata relating them would make up the AIP. As with the public lists, notebook files and their accompanying metadata would make up AICs. When a user retrieves a message, the normalized attachment would be served up as part of the DIP. (See Figure 2.)

Fewer than two thousand of the one million H-Net messages contain attachments of documents and other files of interest, however. At this time, MATRIX has elected to keep attachments embedded in their messages rather than detaching and normalizing them. When a message is retrieved, a link to it will be provided and the user may download it. MATRIX will provide conversion tools or links to websites containing conversion tools for attachments in earlier versions of common formats.⁷

⁶ CCSDS, 4-39 – 4-42.

⁷ Schmidt, “Preservation of the H-Net E-Mail Lists: Suggested Improvements,” 9-11.

Figure 2. H-Net Private List Information Packages (Suggested)

