Internet Engineering Task Force (IETF)
Request for Comments: 6778
Category: Informational
ISSN: 2070-1721

R. Sparks
Tekelec

October 2012

Requirements for Archiving IETF Email Lists and for Providing Web-Based Browsing and Searching

Abstract

The IETF makes heavy use of email lists to conduct its work. Participants frequently need to search and browse the archives of these lists and have asked for improved search capabilities. The current archive mechanism could also be made more efficient. This memo captures the requirements for improved email list archiving and searching systems.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6778.

Table of Contents

1. Introduction

The IETF makes heavy use of email lists to conduct its work. Participants frequently need to search the archives of these lists and have asked for improved search capabilities, particularly when the search needs to cover a large period of time or cross several lists. For instance, document editors, shepherds, working group chairs, and area directors may need to review all discussion of a particular draft. That discussion may be spread across the working group list, one or more directorate lists, and the IETF general list. Occasionally, work impacts multiple groups, possibly in different areas, and the search must cover additional working group lists.

The current tools for performing these searches require several manually coordinated steps, which are error prone. Without a local copy of the archive (which may not be complete), searching most working group lists requires brute force effort, aided possibly by web search engines.

More advanced search capabilities have been constructed for a limited subset of the available lists and are exposed in the "Email Archives Quick Search" section of the main IETF website. While these tools are of great assistance, there is still significant need for improvement.

The current archive mechanism could also be made more efficient. The current practices involve duplicate stores (for the web and ftp interfaces), which impact storage and replication and are subject to inconsistency.

This memo captures the requirements for improved email list archiving and searching systems.

2. List Search and Archive Requirements

2.1. Search and Browsing

- The system must provide a web interface for searching and browsing archived messages.

- The system must allow browsing the entire archive of a given list by thread or by date.

- The system must allow browsing the results of a search by thread or by date.

  Both threading based on Message-Id/References/In-Reply-To and threading based on same subject line (modulo short prefixes like re: and fwd:) should be taken into account.

- The system must allow searching across any subset of the archived lists (one list, a selection of lists, or all lists).

- The system must allow searching of any combination (using AND, OR, and NOT operators) of the following attributes. Richer search capabilities are highly desirable.

  - string occurring in sender name or email address

  - date range

  - string occurring in Subject

  - string occurring in message body

  - string occurring in message header (in particular, exact match of Message-Id)

    For instance, it would be nice to search the entire archive for instances of a message with a given Message-ID with a URL like <http://datatracker.ietf.org/mlarchive/msg?id=4EA6E023.6010603@example.com>

- Individual messages must be representable by a long-term stable URI that can be shared between users. That is, the URI must be suitable for reference in an email message.
  - It would be preferable for that URI to appear in an Archived-At header field in the message [RFC5064].
- Searches should be representable by a URI that can be shared between users.
  - Such URIs should be long-term stable.
  - The search may be re-executed when the URI is referenced. It is acceptable for the same URI to produce different results if accessed at different times or by different people (for example, by reflecting additional messages that may match the search criteria or reflecting changes in access authorization to lists with restricted archives).
- When the system requires credentials, it must use the datatracker's authentication system.
  - While the vast majority of archived lists have an open access policy, some archived lists have restricted archives.
  - The system must not require credentials for browsing or searching lists with open archives. (But it is acceptable for a user to browse or search such lists while logged in.)
  - The system must make it possible to limit access to a restricted archive based on login credentials.
  - Messages from restricted archives must be distinguishable from messages from unrestricted archives in any search results.

2.2. Archiving Active Lists

- The archive system must accept messages handled by various mail list manager packages.
  - Lists hosted on the IETF systems are served by Mailman [Mailman].
  - Lists hosted at other organizations may use other packages.
    - The archive system must accept messages through subscribing to such an external list.
    - The archive system may support other mechanisms for accepting messages into the archive.

2.3. Importing Messages from Other Archives

Lists hosted at other systems are sometimes moved to the IETF servers, and their archive is moved with them. The archiving system must be able to import these archives.

- At a minimum, the archive system must be able to import mbox formatted archives [RFC4155] [Mbox].
- The archive system should be able to import Maildir and Maildir-like (the key characteristic being one-message-per-file) formatted archives [Maildir].
- It is acceptable to use a separate utility to convert between these formats before import, as long as the conversion is lossless.

2.4. Exporting Messages from the Archives

The archive system must allow both users and administrators to export messages.

- The archive system must support exporting messages in the mbox format.
- The archive system should support exporting messages in the Maildir format.
- The archive system must support exporting the entire archive of a given list.

- The archive system must support exporting all messages from a given list within a given date range.

- The archive system should allow exporting the results of any supported search query.

2.5. Redundancy

- The systems must facilitate providing archive, search, and browse functions through geographically distributed servers.

  - The systems must support a single active and single standby server. This reflects the current operating configuration and is expected to be the initial deployment model.

  - The systems should support a single active and multiple standby servers.

  - The systems should support multiple active servers for the search and browse functions. Support for multiple active archive servers is not a requirement.

  - The amount of traffic generated to ensure data replication between servers should be on the order of the size of any new/changed messages in the archives.

    * It is acceptable for replication to be part of the archival system itself (such as using the replication mechanisms from an underlying database).

    * It is acceptable to rely on replication of the underlying filesystem objects (using rsync of one or more directory trees, for example), but only if the objects in the underlying filesystem are formatted such that the size of the replication data is on the order of the size of any new/changed messages in the archives.

2.6. Archive Administration

- The archive system must support adding and removing lists to be archived.

- The system must allow the administrator to add messages to and delete messages from an archived list. The system should log such actions.

2.7. Transition Requirements

There are many existing archived messages containing embedded links into the existing MHonArc [MHonArc] mail archive. These links must continue to work, but they should reach the message as archived in the new system.

3. Internationalized Address Considerations

The archive and search functions should anticipate internationalized email addresses as discussed in the following three documents -- [RFC6531], [RFC6532], and [IMAP-UTF8]. There is no firm requirement at this time.

4. IMAP Access

Requirements for allowing access to the archives using IMAP are captured in [IMAP-ARCH]. The archive system must anticipate integrating with a system that provides IMAP access.

5. Security Considerations

Creating a new tool for searching and archiving IETF email lists does not affect the security of the Internet in any significant fashion.

Searching can be I/O and CPU intensive. The implementors of this tool should consider the potential for maliciously crafted searches attempting to consume all available resources. Similarly, the implementors should consider the potential for denial-of-service attacks through making many connections to the browsing system or rapid navigating within it.

Preserving the integrity of the archives is important. The implementors should ensure that administrative access is appropriately authenticated, and that message paths into the archive are appropriately configured to avoid unauthorized message insertion.

6. Acknowledgements

The Tools Development team provided input into the initial brainstorm. Text suggestions from Alexey Melnikov, Pete Resnick, S. Moonesamy, Francis Dupont, and Murray Kucherawy have been incorporated.

7. Informative References

[IMAP-ARCH] Sparks, R., "IMAP Access to IETF Email List Archives", Work in Progress, August 2012.

[IMAP-UTF8] Resnick, P., Ed., Newman, C., Ed., and S. Shen, Ed., "IMAP Support for UTF-8", Work in Progress, October 2012.

[Maildir] "Maildir", <http://en.wikipedia.org/w/index.php?title=Maildir&oldid=503288041>.

[Mailman] "Mailman", <http://www.list.org/>.

[Mbox] "mbox", <http://en.wikipedia.org/w/index.php?title=Mbox&oldid=515402275>.

[MHonArc] "MHonArc", <http://www.mhonarc.org/>.

[RFC4155] Hall, E., "The application/mbox Media Type", RFC 4155, September 2005.

[RFC5064] Duerst, M., "The Archived-At Message Header Field", RFC 5064, December 2007.

[RFC6531] Yao, J. and W. Mao, "SMTP Extension for Internationalized Email", RFC 6531, February 2012.

[RFC6532] Yang, A., Steele, S., and N. Freed, "Internationalized Email Headers", RFC 6532, February 2012.

Author's Address

Robert Sparks
Tekelec
17210 Campbell Road
Suite 250
Dallas, Texas 75254-4203
USA

EMail: RjS@nostrum.com