Network Working Group
Internet Draft
Intended status: Standards Track
Expires: March 2008
Updates: RFC-ietf-sieve-variables

Jutta Degener
Philip Guenther
Sendmail, Inc.

September 2008

Sieve Email Filtering: Body Extension

draft-ietf-sieve-body-09.txt

Status of this memo

Copyright Notice

Abstract

This document defines a new command for the "Sieve" email filtering language that tests for the occurrence of one or more strings in the body of an email message.

1. Introduction

The "body" test checks for the occurrence of one or more strings in the body of an email message. Such a test was initially discussed for the [SIEVE] base document, but was subsequently removed because it was thought to be too costly to implement.

Nevertheless, several server vendors have implemented some form of the "body" test.

This document reintroduces the "body" test as an extension, and specifies its syntax and semantics.

2. Conventions Used in This Document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [KEYWORDS].

Conventions for notations are as in [SIEVE] section 1.1, including use of the "Usage:" label for the definition of text and tagged arguments syntax.

The rules for interpreting the grammar are defined in [SIEVE] and inherited by this specification. In particular, readers of this document are reminded that according to [SIEVE] sections 2.6.2 and 2.6.3, optional arguments such as COMPARATOR and MATCH-TYPE can appear in any order.

3. Capability Identifier

The capability string associated with the extension defined in this document is "body".

4. Test body

Usage: "body" [COMPARATOR] [MATCH-TYPE] [BODY-TRANSFORM]
<key-list: string-list>

The body test matches content in the body of an email message, that is, anything following the first empty line after the header. (The empty line itself, if present, is not considered to be part of the body.)

The COMPARATOR and MATCH-TYPE keyword parameters are defined in [SIEVE]. As specified in section 2.7.3 of [SIEVE], the default COMPARATOR is "i;ascii-casemap" and the default MATCH-TYPE is ":is".

The BODY-TRANSFORM is a keyword parameter that governs how a set of strings to be matched against are extracted from the body of the message. If a message consists of a header only, not followed by an empty line, then that set is empty and all "body" tests return false, including those that test for an empty string. (This is similar to how the "header" test always fails when the named header fields aren't present.) Otherwise, the transform must be followed as defined below in section 4.

Note that the transforms defined here do *not* match against each line of the message independently, so the strings will usually contain CRLFs. How these can be matched is governed by the comparator and match-type. For example, with the default comparator of "i;ascii-casemap", they can be included literally in key string, or be matched with the "*" or "?" wildcards of the :matches match-type, or be skipped with :contains.

5. Body Transform

Prior to matching content in a message body, "transformations" can be applied that filter and decode certain parts of the body. These transformations are selected by a "BODY-TRANSFORM" keyword parameter.

```
Usage: ":raw"
     / ":content" <content-types: string-list>
     / ":text"
```

The default transformation is :text.

5.1 Body Transform ":raw"

The ":raw" transform matches against the entire, undecoded body of a message as a single item.

If the specified body-transform is ":raw", the [MIME] structure of the body is irrelevant. The implementation MUST NOT remove any transfer encoding from the message, MUST NOT refuse to filter messages with syntactic errors (unless the environment it is part of rejects them outright), and MUST treat multipart boundaries or the MIME headers of enclosed body parts as part of the content being matched against instead of MIME structures to interpret.

Example:

```
require "body";

# This will match a message containing the literal text
# "MAKE MONEY FAST" in body parts (ignoring any
# content-transfer-encodings) or MIME headers other than
# the outermost RFC 2822 header.

if body :raw :contains "MAKE MONEY FAST" {
        discard;
}
```

5.2 Body Transform ":content"

If the body transform is ":content", the MIME parts that have the specified content-types are matched against independently, each the entire part as a single string.

If an individual content type begins or ends with a '/' (slash) or contains multiple slashes, it matches no content types. Otherwise, if it contains a slash, then it specifies a full <type>/<subtype> pair, and matches only that specific content type. If it is the empty string, all MIME content types are matched. Otherwise, it specifies a <type> only, and any subtype of that type matches it.

The search for MIME parts matching the :content specification is recursive and automatically descends into multipart and message/rfc822 MIME parts. All MIME parts with matching types are searched for the key strings. The test returns true if any combination of searched MIME part and key-list argument match.

If the :content specification matches a multipart MIME part, only the prologue and epilogue sections of the part will be searched for the key strings; the contents of nested parts are only searched if their respective types match the :content specification.

If the :content specification matches a message/rfc822 MIME part, only the header of the nested message will be searched for the key strings; the contents of the nested message body parts are only searched if its content-type matches the :content specification.

(Matches against container types with an empty match string can be useful as tests for the existence of such parts.)

```
Example:
     From: Whomever
     To: Someone
     Date: Whenever
     Subject: whatever
     Content-Type: multipart/mixed; boundary=outer

  &  This is a multi-part message in MIME format.
  &
     --outer
     Content-Type: multipart/alternative; boundary=inner

  &  This is a nested multi-part message in MIME format.
  &
     --inner
     Content-Type: text/plain; charset="us-ascii"

  $  Hello
  $
     --inner
     Content-Type: text/html; charset="us-ascii"

  %  <html><body>Hello</body></html>
  %
     --inner--
  &
  &  This is the end of the inner MIME multipart.
  &
     --outer
     Content-Type: message/rfc822

  !  From: Someone Else
  !  Subject: hello request

  $  Please say Hello
  $
     --outer--
  &
  &  This is the end of the outer MIME multipart.
```

In the above example, the '&', '$', '%', and '!' characters at the start of a line are used to illustrate what portions of the example message are used in tests:

- the lines starting with '&' are the ones that are tested when a 'body :content "multipart" :contains "MIME"' test is executed.

- the lines starting with '$' are the ones that are tested when a 'body :content "text/plain" :contains "Hello"' test is executed.

- the lines starting with '%' are the ones that are tested when a 'body :content "text/html" :contains "Hello"' test is executed.

- the lines starting with '$' or '%' are the ones that are tested when a 'body :content "text" :contains "Hello"' test is executed.

- the lines starting with '!' are the ones that are tested when a 'body :content "message/rfc822" :contains "Hello"' test is executed.

Comparisons are performed on octets. Implementations decode the content-transfer-encoding and convert text to [UTF-8] as input to the comparator. MIME parts that cannot be decoded and converted MAY be treated as plain US-ASCII, omitted, or processed according to local conventions. A NUL octet (character zero) SHOULD NOT cause early termination of the content being compared against. Implementations MUST support the "quoted-printable", "base64", "7bit", "8bit", and "binary" content transfer encodings. Implementations MUST be capable of converting to UTF-8 the US-ASCII, ISO-8859-1, and the US-ASCII subset of ISO-8859-* character sets.

Each matched part is matched against independently: search expressions MUST NOT match across MIME part boundaries. MIME headers of the containing part MUST NOT be included in the data.

Example:

```
     require ["body", "fileinto"];

     # Save any message with any text MIME part that contains the
     # words "missile" or "coordinates" in the "secrets" folder.

     if body :content "text" :contains ["missile", "coordinates"] {
             fileinto "secrets";
     }

     # Save any message with an audio/mp3 MIME part in
     # the "jukebox" folder.

     if body :content "audio/mp3" :contains "" {
             fileinto "jukebox";
     }
```

5.3 Body Transform ":text"

The ":text" body transform matches against the results of an implementation's best effort at extracting UTF-8 encoded text from a message.

It is unspecified whether this transformation results in a single string or multiple strings being matched against. All the text extracted from a given non-container MIME part MUST be in the same string

In simple implementations, :text MAY be treated the same as :content "text".

Sophisticated implementations MAY strip mark-up from the text prior to matching, and MAY convert media types other than text to text prior to matching.

(For example, they may be able to convert proprietary text editor formats to text or apply optical character recognition algorithms to image data.)

Example:

```
require ["body", "fileinto"];

# Save messages mentioning the project schedule in the
# project/schedule folder.
if body :text :contains "project schedule" {
        fileinto "project/schedule";
}
```

6. Interaction with Other Sieve Extensions

Any extension that extends the grammar for the COMPARATOR or MATCH-TYPE nonterminals will also affect the implementation of "body".

Wildcard expressions used with "body" are exempt from the side effects described in [VARIABLES]. That is, they MUST NOT set match variables (${1}, ${2}...) to the input values corresponding to wildcard sequences in the matched pattern. However, if the extension is present, variable references in the key strings or content type strings are evaluated as described in the draft.

7. IANA Considerations

The following template specifies the IANA registration of the Sieve extension specified in this document:

To: iana@iana.org
Subject: Registration of new Sieve extension

Capability name: body
Description: Provides a test for matching against the the body of the message being processed
RFC number: this RFC
Contact Address: Jutta Degener <jutta@pobox.com>

This information should be added to the list of sieve extensions given on http://www.iana.org/assignments/sieve-extensions.

8. Security Considerations

The system MUST be sized and restricted in such a manner that even malicious use of body matching does not deny service to other users of the host system.

Filters relying on string matches in the raw body of an email message may be more general than intended. Text matches are no replacement for a spam, virus, or other security related filtering system.

9. Acknowledgments

This document has been revised in part based on comments and discussions that took place on and off the SIEVE mailing list. Thanks to Cyrus Daboo, Ned Freed, Bob Johannessen, Simon Josefsson, Mark E. Mallett, Chris Markle, Alexey Melnikov, Ken Murchison, Greg Shapiro, Tim Showalter, Nigel Swinson, Dowson Tong, and Christian Vogt for reviews and suggestions.

10. Authors' Addresses

Jutta Degener
5245 College Ave, Suite #127
Oakland, CA 94618

Email: jutta@pobox.com

Philip Guenther
Sendmail, Inc.
6425 Christie Ave, 4th Floor
Emeryville, CA 94608

Email: guenther@sendmail.com

11. Discussion

This section will be removed when this document leaves the Internet-Draft stage.

This draft is intended as an extension to the Sieve mail filtering language. Sieve extensions are discussed on the MTA Filters mailing list at <ietf-mta-filters@imc.org>. Subscription requests can be sent to <ietf-mta-filters-request@imc.org> (send an email message with the word "subscribe" in the body).

More information on the mailing list along with a WWW archive of back messages is available at <http://www.imc.org/ietf-mta-filters/>.

11.1 Changes from draft-ietf-sieve-body-08.txt

Add a "Capability Identifier" section to match existing RFCs.

Make the normative and information references subsections of a "References" section to match existing RFCs.

Tweak description field of the IANA registration.

Change "wild card" to "wildcard".

11.2 Changes from draft-ietf-sieve-body-07.txt

Clarify how transforms generate one or more strings to match against.

Reiterate the default COMPARATOR and MATCH-TYPE from the base spec.

[SIEVE] and [VARIABLES] have been published.

11.3 Changes from draft-ietf-sieve-body-06.txt

Changed "matched text" to "matched content". Drop the word "proposed".

11.4 Changes from draft-ietf-sieve-body-05.txt

Updated boilerplate to match RFC 4748.

Added "Intended-Status: Standards Track" and "Updates: draft-ietf-sieve-variables-08"

Change the references from appendices to sections. Update [SIEVE] reference.

11.5 Changes from draft-ietf-sieve-body-04.txt

Changed 'reject' to 'discard' in the example.

Removed reference to regex draft.

Update copyright boilerplate.

11.6 Changes from draft-ietf-sieve-body-03.txt

Update IANA registration to match 3028bis.

Added direct boilerplate for [KEYWORDS].

11.7 Changes from draft-ietf-sieve-body-02.txt

Updated charset conversion to match draft-ietf-sieve-3028bis-06.txt.

Change "Syntax:" to "Usage:".

Updated references.

11.8 Changes from draft-ietf-sieve-body-01.txt

Updated charset conversion requirements to match those in draft-ietf-sieve-3028bis-03.txt for headers.

11.9 Changes from draft-ietf-sieve-body-00.txt

Updated IPR boilerplate to RFC 3978/3979.

Many prose corrections in response to WGLC comments. Of particular note:

- made clear that :raw treats MIME boundaries and headers as text to be matched against
- corrected description in comment of :raw example
- clarified the interpretation of invalid content-types in :content
- gave precise description of what gets matched when :content is used with message/rfc822 or any multipart type, as well as a comprehensive example
- include an example of :text
- tightened wording of interaction with [VARIABLES]
- added informative reference to [REGEX]

11.10 Changes from draft-degener-sieve-body-04.txt

Renamed to draft-ietf-sieve-body-00.txt; tweaked the title and abstract.

Added Philip Guenther as co-author.

Split references into normative and informative. Updated [UTF-8] and [VARIABLES] references.

Updated IPR boilerplate.

11.11 Changes from draft-degener-sieve-body-03.txt

Made "body" exempt from variable-setting side effects in the presence of the "variables" extension and wildcards. It's too hard to implement.

Removed :binary. It's uglier and less useful than it needs to be to bother.

Added IANA section.

12. References

12.1. Normative References

[KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 15, RFC 2119, March 1997.

[MIME] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies", RFC 2045, November 1996.

[SIEVE] Guenther, P. and T. Showalter, "Sieve: An Email Filtering Language", RFC 5228, January 2008.

[UTF-8] Yergeau, F., "UTF-8, a transformation format of ISO 10646", RFC 3629, November 2003.

12.2. Informative References

[VARIABLES] Homme, K., "Sieve Email Filtering: Variables Extension", RFC 5229, January 2008.

Full Copyright Statement

Intellectual Property

Acknowledgement

Funding for the RFC Editor function is currently provided by the IETF Administrative Support Activity (IASA).