Network Working Group
Request for Comments: 5738
Updates: 3501
Category: Experimental

P. Resnick
Qualcomm Incorporated
C. Newman
Sun Microsystems

March 2010

IMAP Support for UTF-8

Abstract

This specification extends the Internet Message Access Protocol version 4rev1 (IMAP4rev1) to support UTF-8 encoded international characters in user names, mail addresses, and message headers.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for examination, experimental implementation, and evaluation.

This document defines an Experimental Protocol for the Internet community. 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/rfc5738.

Table of Contents

1. Introduction

This specification extends IMAP4rev1 [RFC3501] to permit UTF-8 [RFC3629] in headers as described in "Internationalized Email Headers" [RFC5335]. It also adds a mechanism to support mailbox names, login names, and passwords using the UTF-8 charset. This specification creates five new IMAP capabilities to allow servers to advertise these new extensions, along with two new IMAP LIST selection options and a new IMAP LIST return option.

2. Conventions Used in This Document

The key words "MUST", "MUST NOT", "SHOULD", "SHOULD NOT", and "MAY" in this document are to be interpreted as defined in "Key words for use in RFCs to Indicate Requirement Levels" [RFC2119].

The formal syntax uses the Augmented Backus-Naur Form (ABNF) [RFC5234] notation including the core rules defined in Appendix B of [RFC5234]. In addition, rules from IMAP4rev1 [RFC3501], UTF-8 [RFC3629], "Collected Extensions to IMAP4 ABNF" [RFC4466], and IMAP4 LIST Command Extensions [RFC5258] are also referenced.

In examples, "C:" and "S:" indicate lines sent by the client and server, respectively. If a single "C:" or "S:" label applies to multiple lines, then the line breaks between those lines are for editorial clarity only and are not part of the actual protocol exchange.

3. UTF8=ACCEPT IMAP Capability

The "UTF8=ACCEPT" capability indicates that the server supports UTF-8 quoted strings, the "UTF8" parameter to SELECT and EXAMINE, and UTF-8 responses from the LIST and LSUB commands.

A client MUST use the "ENABLE UTF8=ACCEPT" command (defined in [RFC5161]) to indicate to the server that the client accepts UTF-8 quoted-strings. The "ENABLE UTF8=ACCEPT" command MUST only be used in the authenticated state. (Note that the "UTF8=ONLY" capability described in Section 7 and the "UTF8=ALL" capability described in Section 6 imply the "UTF8=ACCEPT" capability. See additional information in these sections.)

3.1. IMAP UTF-8 Quoted Strings

The IMAP4rev1 [RFC3501] base specification forbids the use of 8-bit characters in atoms or quoted strings. Thus, a UTF-8 string can only be sent as a literal. This can be inconvenient from a coding standpoint, and unless the server offers IMAP4 non-synchronizing

literals [RFC2088], this requires an extra round trip for each UTF-8 string sent by the client. When the IMAP server advertises the "UTF8=ACCEPT" capability, it informs the client that it supports native UTF-8 quoted-strings with the following syntax:

```
string        =/ utf8-quoted

utf8-quoted   = "*" DQUOTE *UQUOTED-CHAR DQUOTE

UQUOTED-CHAR  = QUOTED-CHAR / UTF8-2 / UTF8-3 / UTF8-4
           ; UTF8-2, UTF8-3, and UTF8-4 are as defined in RFC 3629
```

When this quoting mechanism is used by the client (specifically an octet sequence beginning with *" and ending with "), then the server MUST reject octet sequences with the high bit set that fail to comply with the formal syntax in [RFC3629] with a BAD response.

The IMAP server MUST NOT send utf8-quoted syntax to the client unless the client has indicated support for that syntax by using the "ENABLE UTF8=ACCEPT" command.

If the server advertises the "UTF8=ACCEPT" capability, the client MAY use utf8-quoted syntax with any IMAP argument that permits a string (including astring and nstring). However, if characters outside the US-ASCII repertoire are used in an inappropriate place, the results would be the same as if other syntactically valid but semantically invalid characters were used. For example, if the client includes UTF-8 characters in the user or password arguments (and the server has not advertised "UTF8=USER"), the LOGIN command will fail as it would with any other invalid user name or password. Specific cases where UTF-8 characters are permitted or not permitted are described in the following paragraphs.

All IMAP servers that advertise the "UTF8=ACCEPT" capability SHOULD accept UTF-8 in mailbox names, and those that also support the "Mailbox International Naming Convention" described in RFC 3501, Section 5.1.3 MUST accept utf8-quoted mailbox names and convert them to the appropriate internal format. Mailbox names MUST comply with the Net-Unicode Definition (Section 2 of [RFC5198]) with the specific exception that they MUST NOT contain control characters (0000-001F, 0080-009F), delete (007F), line separator (2028), or paragraph separator (2029).

An IMAP client MUST NOT issue a SEARCH command that uses a mixture of utf8-quoted syntax and a SEARCH CHARSET other than UTF-8. If an IMAP server receives such a SEARCH command, it SHOULD reject the command with a BAD response (due to the conflicting charset labels).

3.2. UTF8 Parameter to SELECT and EXAMINE

The "UTF8=ACCEPT" capability also indicates that the server supports the "UTF8" parameter to SELECT and EXAMINE. When a mailbox is selected with the "UTF8" parameter, it alters the behavior of all IMAP commands related to message sizes, message headers, and MIME body headers so they refer to the message with UTF-8 headers. If the mailstore is not UTF-8 header native and the SELECT or EXAMINE command with UTF-8 header modifier succeeds, then the server MUST return results as if the mailstore were UTF-8 header native with upconversion requirements as described in Section 8. The server MAY reject the SELECT or EXAMINE command with the [NOT-UTF-8] response code, unless the "UTF8=ALL" or "UTF8=ONLY" capability is advertised.

Servers MAY include mailboxes that can only be selected or examined if the "UTF8" parameter is provided. However, such mailboxes MUST NOT be included in the output of an unextended LIST, LSUB, or equivalent command. If a client attempts to SELECT or EXAMINE such mailboxes without the "UTF8" parameter, the server MUST reject the command with a [UTF-8-ONLY] response code. As a result, such mailboxes will not be accessible by IMAP clients written prior to this specification and are discouraged unless the server advertises "UTF8=ONLY" or the server implements IMAP4 LIST Command Extensions [RFC5258].

```
utf8-select-param = "UTF8"
          ;; Conforms to <select-param> from RFC 4466
```

```
C: a SELECT newmailbox (UTF8)
S: ...
S: a OK SELECT completed
C: b FETCH 1 (SIZE ENVELOPE BODY)
S: ... < UTF-8 header native results >
S: b OK FETCH completed

C: c EXAMINE legacymailbox (UTF8)
S: c NO [NOT-UTF-8] Mailbox does not support UTF-8 access

C: d SELECT funky-new-mailbox
S: d NO [UTF-8-ONLY] Mailbox requires UTF-8 client
```

3.3. UTF-8 LIST and LSUB Responses

After an IMAP client successfully issues an "ENABLE UTF8=ACCEPT" command, the server MUST NOT return in LIST results any mailbox names to the client following the IMAP4 Mailbox International Naming Convention. Instead, the server MUST return any mailbox names with characters outside the US-ASCII repertoire using utf8-quoted syntax.

(The IMAP4 Mailbox International Naming Convention has proved problematic in the past, so the desire is to make this syntax obsolete as quickly as possible.)

3.4. UTF-8 Interaction with IMAP4 LIST Command Extensions

When an IMAP server advertises both the "UTF8=ACCEPT" capability and the "LIST-EXTENDED" [RFC5258] capability, the server MUST support the LIST extensions described in this section.

3.4.1. UTF8 and UTF8ONLY LIST Selection Options

The "UTF8" LIST selection option tells the server to include mailboxes that only support UTF-8 headers in the output of the list command. The "UTF8ONLY" LIST selection option tells the server to include all mailboxes that support UTF-8 headers and to exclude mailboxes that don't support UTF-8 headers. Note that "UTF8ONLY" implies "UTF8", so it is not necessary for the client to request both. Use of either selection option will also result in UTF-8 mailbox names in the result as described in Section 3.3 and implies the "UTF8" List return option described in Section 3.4.2.

3.4.2. UTF8 LIST Return Option

If the client supplies the "UTF8" LIST return option, then the server MUST include either the "\NoUTF8" or the "\UTF8Only" mailbox attribute as appropriate. The "\NoUTF8" mailbox attribute indicates that an attempt to SELECT or EXAMINE that mailbox with the "UTF8" parameter will fail with a [NOT-UTF-8] response code. The "\UTF8Only" mailbox attribute indicates that an attempt to SELECT or EXAMINE that mailbox without the "UTF8" parameter will fail with a [UTF-8-ONLY] response code. Note that computing this information may be expensive on some server implementations, so this return option should not be used unless necessary.

The ABNF [RFC5234] for these LIST extensions follows:

```
  list-select-independent-opt =/ "UTF8"

  list-select-base-opt        =/ "UTF8ONLY"

  mbx-list-oflag              =/ "\NoUTF8" / "\UTF8Only"

  return-option               =/ "UTF8"

  resp-text-code              =/ "NOT-UTF-8" / "UTF-8-ONLY"
```

4. UTF8=APPEND Capability

If the "UTF8=APPEND" capability is advertised, then the server accepts UTF-8 headers in the APPEND command message argument. A client that sends a message with UTF-8 headers to the server MUST send them using the "UTF8" APPEND data extension. If the server also advertises the CATENATE capability (as specified in [RFC4469]), the client can use the same data extension to include such a message in a CATENATE message part. The ABNF for the APPEND data extension and CATENATE extension follows:

```
utf8-literal   = "UTF8" SP "(" literal8 ")"

append-data    =/ utf8-literal

cat-part       =/ utf8-literal
```

A server that advertises "UTF8=APPEND" has to comply with the requirements of the IMAP base specification and [RFC5322] for message fetching. Mechanisms for 7-bit downgrading to help comply with the standards are discussed in Downgrading mechanism for Internationalized eMail Address (IMA) [RFC5504].

IMAP servers that do not advertise the "UTF8=APPEND" or "UTF8=ONLY" capability SHOULD reject an APPEND command that includes any 8-bit in the message headers with a "NO" response.

Note that the "UTF8=ONLY" capability described in Section 7 implies the "UTF8=APPEND" capability. See additional information in that section.

5. UTF8=USER Capability

If the "UTF8=USER" capability is advertised, that indicates the server accepts UTF-8 user names and passwords and applies SASLprep [RFC4013] to both arguments of the LOGIN command. The server MUST reject UTF-8 that fails to comply with the formal syntax in RFC 3629 [RFC3629] or if it encounters Unicode characters listed in Section 2.3 of SASLprep RFC 4013 [RFC4013].

6. UTF8=ALL Capability

The "UTF8=ALL" capability indicates all server mailboxes support UTF-8 headers. Specifically, SELECT and EXAMINE with the "UTF8" parameter will never fail with a [NOT-UTF-8] response code.

Note that the "UTF8=ONLY" capability described in Section 7 implies the "UTF8=ALL" capability. See additional information in that section.

Note that the "UTF8=ALL" capability implies the "UTF8=ACCEPT" capability.

7. UTF8=ONLY Capability

The "UTF8=ONLY" capability permits an IMAP server to advertise that it does not support the international mailbox name convention (modified UTF-7), and does not permit selection or examination of any mailbox unless the "UTF8" parameter is provided. As this is an incompatible change to IMAP, a clear warning is necessary. IMAP clients that find implementation of the "UTF8=ONLY" capability problematic are encouraged to at least detect the "UTF8=ONLY" capability and provide an informative error message to the end-user.

When an IMAP mailbox internally uses UTF-8 header native storage, the down-conversion step is necessary to permit selection or examination of the mailbox in a backwards compatible fashion will become more difficult to support. Although it is hoped that deployed IMAP servers will not advertise "UTF8=ONLY" for some years, this capability is intended to minimize the disruption when legacy support finally goes away.

The "UTF8=ONLY" capability implies the "UTF8=ACCEPT" capability, the "UTF8=ALL" capability, and the "UTF8=APPEND" capability. A server that advertises "UTF8=ONLY" need not advertise the three implicit capabilities.

8. Up-Conversion Server Requirements

When an IMAP4 server uses a traditional mailbox format that includes 7-bit headers and it chooses to permit access to that mailbox with the "UTF8" parameter, it MUST support minimal up-conversion as described in this section.

The server MUST support up-conversion of the following address header-fields in the message header: From, Sender, To, CC, Bcc, Resent-From, Resent-Sender, Resent-To, Resent-CC, Resent-Bcc, and Reply-To. This up-conversion MUST include address local-parts in fields downgraded according to [RFC5504], address domains encoded according to Internationalizing Domain Names in Applications (IDNA) [RFC3490], and MIME header encoding [RFC2047] of display-names and any [RFC5322] comments.

The following charsets MUST be supported for up-conversion of MIME header encoding [RFC2047]: UTF-8, US-ASCII, ISO-8859-1, ISO-8859-2, ISO-8859-3, ISO-8859-4, ISO-8859-5, ISO-8859-6, ISO-8859-7, ISO-8859-8, ISO-8859-9, ISO-8859-10, ISO-8859-14, and ISO-8859-15. If the server supports other charsets in IMAP SEARCH or IMAP CONVERT [RFC5259], it SHOULD also support those charsets in this conversion.

Up-conversion of MIME header encoding of the following headers MUST also be implemented: Subject, Date ([RFC5322] comments only), Comments, Keywords, and Content-Description.

Server implementations also SHOULD up-convert all MIME body headers [RFC2045], SHOULD up-convert or remove the deprecated (and misused) "name" parameter [RFC1341] on Content-Type, and MUST up-convert the Content-Disposition [RFC2183] "filename" parameter, except when any of these are contained within a multipart/signed MIME body part (see below). These parameters can be encoded using the standard MIME parameter encoding [RFC2231] mechanism, or via non-standard use of MIME header encoding [RFC2047] in quoted strings.

The IMAP server MUST NOT perform up-conversion of headers and content of multipart/signed, as well as Original-Recipient and Return-Path.

9. Issues with UTF-8 Header Mailstore

When an IMAP server uses a mailbox format that supports UTF-8 headers and it permits selection or examination of that mailbox without the "UTF8" parameter, it is the responsibility of the server to comply with the IMAP4rev1 base specification [RFC3501] and [RFC5322] with respect to all header information transmitted over the wire. Mechanisms for 7-bit downgrading to help comply with the standards are discussed in "Downgrading Mechanism for Email Address Internationalization" [RFC5504].

An IMAP server with a mailbox that supports UTF-8 headers MUST comply with the protocol requirements implicit from Section 8. However, the code necessary for such compliance need not be part of the IMAP server itself in this case. For example, the minimal required up-conversion could be performed when a message is inserted into the IMAP-accessible mailbox.

10. IANA Considerations

This adds five new capabilities ("UTF8=ACCEPT", "UTF8=USER", "UTF8=APPEND", "UTF8=ALL", and "UTF8=ONLY") to the IMAP4rev1 Capabilities registry [RFC3501].

This adds two new IMAP4 list selection options and one new IMAP4 list return option.

1. LIST-EXTENDED option name: UTF8

   LIST-EXTENDED option type: SELECTION

   Implied return options(s): UTF8

   LIST-EXTENDED option description: Causes the LIST response to include mailboxes that mandate the UTF8 SELECT/EXAMINE parameter.

   Published specification: RFC 5738, Section 3.4.1

   Security considerations: RFC 5738, Section 11

   Intended usage: COMMON

   Person and email address to contact for further information: see the Authors' Addresses at the end of this specification

   Owner/Change controller: iesg@ietf.org

2. LIST-EXTENDED option name: UTF8ONLY

   LIST-EXTENDED option type: SELECTION

   Implied return options(s): UTF8

   LIST-EXTENDED option description: Causes the LIST response to include mailboxes that mandate the UTF8 SELECT/EXAMINE parameter and exclude mailboxes that do not support the UTF8 SELECT/EXAMINE parameter.

   Published specification: RFC 5738, Section 3.4.1

   Security considerations: RFC 5738, Section 11

   Intended usage: COMMON

   Person and email address to contact for further information: see the Authors' Addresses at the end of this specification

   Owner/Change controller: iesg@ietf.org

3. LIST-EXTENDED option name: UTF8

   LIST-EXTENDED option type: RETURN

   Implied return options(s): none

   LIST-EXTENDED option description: Causes the LIST response to include \NoUTF8 and \UTF8Only mailbox attributes.

   Published specification: RFC 5738, Section 3.4.1

   Security considerations: RFC 5738, Section 11

   Intended usage: COMMON

   Person and email address to contact for further information: see the Authors' Addresses at the end of this specification

   Owner/Change controller: iesg@ietf.org

11. Security Considerations

The security considerations of UTF-8 [RFC3629] and SASLprep [RFC4013] apply to this specification, particularly with respect to use of UTF-8 in user names and passwords. Otherwise, this is not believed to alter the security considerations of IMAP4rev1.

12. References

12.1. Normative References

[RFC1341] Borenstein, N. and N. Freed, "MIME (Multipurpose Internet Mail Extensions): Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1341, June 1992.

[RFC2045] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies", RFC 2045, November 1996.

[RFC2047] Moore, K., "MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text", RFC 2047, November 1996.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2183] Troost, R., Dorner, S., and K. Moore, "Communicating Presentation Information in Internet Messages: The Content-Disposition Header Field", RFC 2183, August 1997.

[RFC2231] Freed, N. and K. Moore, "MIME Parameter Value and Encoded Word Extensions:
Character Sets, Languages, and Continuations", RFC 2231, November 1997.

[RFC3490] Faltstrom, P., Hoffman, P., and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, March 2003.

[RFC3501] Crispin, M., "INTERNET MESSAGE ACCESS PROTOCOL - VERSION 4rev1", RFC 3501, March 2003.

[RFC3629] Yergeau, F., "UTF-8, a transformation format of ISO 10646", STD 63, RFC 3629, November 2003.

[RFC4013] Zeilenga, K., "SASLprep: Stringprep Profile for User Names and Passwords", RFC 4013, February 2005.

[RFC4466] Melnikov, A. and C. Daboo, "Collected Extensions to IMAP4 ABNF", RFC 4466, April 2006.

[RFC4469] Resnick, P., "Internet Message Access Protocol (IMAP) CATENATE Extension", RFC 4469, April 2006.

[RFC5161] Gulbrandsen, A. and A. Melnikov, "The IMAP ENABLE Extension", RFC 5161, March 2008.

[RFC5198] Klensin, J. and M. Padlipsky, "Unicode Format for Network Interchange", RFC 5198, March 2008.

[RFC5234] Crocker, D. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, January 2008.

[RFC5258] Leiba, B. and A. Melnikov, "Internet Message Access Protocol version 4 - LIST Command Extensions", RFC 5258, June 2008.

[RFC5259] Melnikov, A. and P. Coates, "Internet Message Access Protocol - CONVERT Extension", RFC 5259, July 2008.

[RFC5322] Resnick, P., Ed., "Internet Message Format", RFC 5322, October 2008.

[RFC5335] Abel, Y., "Internationalized Email Headers", RFC 5335, September 2008.

[RFC5504] Fujiwara, K. and Y. Yoneya, "Downgrading Mechanism for Email Address Internationalization", RFC 5504, March 2009.

12.2. Informative References

[RFC2049] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples", RFC 2049, November 1996.

[RFC2088] Myers, J., "IMAP4 non-synchronizing literals", RFC 2088, January 1997.

[RFC2277] Alvestrand, H., "IETF Policy on Character Sets and Languages", BCP 18, RFC 2277, January 1998.

[RFC5721] Gellens, R. and C. Newman, "POP3 Support for UTF-8", RFC 5721, February 2010.

Appendix A. Design Rationale

This non-normative section discusses the reasons behind some of the design choices in the above specification.

The basic approach of advertising the ability to access a mailbox in UTF-8 mode is intended to permit graceful upgrade, including servers that support multiple mailbox formats. In particular, it would be undesirable to force conversion of an entire server mailstore to UTF-8 headers, so being able to phase-in support for new mailboxes and gradually migrate old mailboxes is permitted by this design.

"UTF8=USER" is optional because many identity systems are US-ASCII only, so it's helpful to inform the client up front that UTF-8 won't work.

"UTF8=APPEND" is optional because it effectively requires IMAP server support for down-conversion, which is a much more complex operation than up-conversion.

The "UTF8=ONLY" mechanism simplifies diagnosis of interoperability problems when legacy support goes away. In the situation where backwards compatibility is broken anyway, just-send-UTF-8 IMAP has the advantage that it might work with some legacy clients. However, the difficulty of diagnosing interoperability problems caused by a just-send-UTF-8 IMAP mechanism is the reason the "UTF8=ONLY" capability mechanism was chosen.

The up-conversion requirements are designed to balance the desire to deprecate and eventually eliminate complicated encodings (like MIME header encodings) without creating a significant deployment burden for servers. As IMAP4 servers already require a MIME parser, this includes additional server up-conversion requirements not present in POP3 Support for UTF-8 [RFC5721].

The set of mandatory charsets comes from two sources: MIME requirements [RFC2049] and IETF Policy on Character Sets [RFC2277]. Including a requirement to up-convert widely deployed encoded ideographic charsets to UTF-8 would be reasonable for most scenarios, but may require unacceptable table sizes for some embedded devices. The open-ended recommendation to support widely deployed charsets avoids the political ramifications of attempting to list such charsets. The authors believe market forces, existing open-source software, and public conversion tables are sufficient to deploy the appropriate charsets.

Appendix B. Examples Demonstrating Relationships between UTF8= Capabilities

```
UTF8=ACCEPT UTF8=USER UTF8=APPEND
UTF8=ACCEPT UTF8=ALL
UTF8=ALL       ; Note, same as above
UTF8=ACCEPT UTF8=USER UTF8=APPEND UTF8=ALL UTF8=ONLY
UTF8=USER UTF8=ONLY ; Note, same as above
```

Appendix C. Acknowledgments

The authors wish to thank the participants of the EAI working group for their contributions to this document with particular thanks to Harald Alvestrand, David Black, Randall Gellens, Arnt Gulbrandsen, Kari Hurtta, John Klensin, Xiaodong Lee, Charles Lindsey, Alexey Melnikov, Subramanian Moonesamy, Shawn Steele, Daniel Taharlev, and Joseph Yee for their specific contributions to the discussion.

Authors' Addresses

Pete Resnick
Qualcomm Incorporated
5775 Morehouse Drive
San Diego, CA 92121-1714
US

Phone: +1 858 651 4478
EMail: presnick@qualcomm.com
URI: http://www.qualcomm.com/~presnick/

Chris Newman
Sun Microsystems
800 Royal Oaks
Monrovia, CA 91016
US

EMail: chris.newman@sun.com