IMAP Support for UTF-8
draft-ietf-eai-imap-utf8-00

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on November 27, 2006.

Copyright Notice

Copyright (C) The Internet Society (2006).

Abstract

This specification extends the Internet Message Access Protocol version 4rev1 (IMAP4rev1) to support unencoded international characters in user names, mail addresses and message headers. This is an early draft and intended as a framework for discussion. Please do not deploy implementations of this draft.
1. 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 used the Augmented Backus-Naur Form (ABNF) [RFC4234] notation including the core rules defined in Appendix B of RFC 4234. In addition, rules from IMAP4rev1 [RFC3501], UTF-8 [RFC3629], Collected extensions to IMAP4 ABNF [I-D.ietf-imapext-list-extensions], and IMAP4 LIST Command Extensions [I-D.ietf-imapext-list-extensions] 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.

2. Introduction

This specification extends IMAP4rev1 [RFC3501] to permit unencoded UTF-8 [RFC3629] in headers as described in Transmission of Email Headers in UTF-8 Encoding [I-D.yeh-ima-utf8headers]. It also adds a mechanism to support mailbox names, login names and passwords using the UTF-8 charset.

3. UTF8 IMAP Capability

The basic "UTF8" capability indicates the server supports UTF-8 quoted strings and the UTF8 parameter to SELECT and EXAMINE.

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" capability, it informs the client that it supports native UTF-8 quoted-strings with the following syntax:
Internet-Draft IMAP Support for UTF-8 May 2006

```
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 which fail to comply with the formal syntax in RFC 3629 [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, either by using it or by using the UTF8 parameter to SELECT or EXAMINE described in Section 3.2.

If the UTF8 capability is advertised, then utf8-quoted syntax MAY be used with any IMAP argument that permits a string or an astring. 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 SHOULD accept UTF-8 in mailbox names and IMAP servers which 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. [TBD stringprep for mailbox names? Can we reuse SASLprep?]

IMAP servers MUST NOT accept UTF-8 characters when storing a new message keyword, unless the mailbox is UTF-8 only, in which case IMAP servers SHOULD accept UTF-8 in message keywords. [TBD stringprep for message keywords? Can we reuse SASLprep?]

If an IMAP client issues a SEARCH command which uses a mixture of utf8-quoted syntax and a SEARCH CHARSET other than UTF-8, then the IMAP server SHOULD reject the command with a BAD response (due to the conflicting charset labels).
3.2. UTF8 Parameter to SELECT and EXAMINE

The "UTF8" capability also indicates the server supports the UTF8 parameter to SELECT and EXAMINE. When a mailbox is selected with the UTF8 parameter, that 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 was 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 which 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 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 [I-D.ietf-imapext-list-extensions].

TBD: describe syntax based on draft-melnikov-imap-ext-abnf-05.

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: s NO [UTF-8-ONLY] Mailbox requires UTF-8 client

3.3. UTF-8 LIST and LSUB Responses

If an IMAP client has used utf8-quoted syntax prior to the server’s production of LIST results, then the server MUST NOT return 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. An IMAP client can force this behavior by issuing the LIST

or LSUB command as follows:

C: a LIST "*" %
S: <responses with UTF-8 mailbox names>
C: a LSUB "*"
S: <responses with UTF-8 mailbox names>
S: a OK LSUB completed

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" capability and the "LIST-EXTENDED" [I-D.ietf-imapext-list-extensions] capability, the server MUST support the LIST extensions described in this section. When an IMAP server advertises the UTF8=ONLY capability and the LIST-EXTENDED capability, the server MUST reject these LIST extensions with a BAD response.

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.

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 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 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 [RFC4234] for these LIST extensions follows:

\[
\text{list-select-independent-opt } =\!/ \text{ "UTF8" / "UTF8ONLY"}
\]

\[
\text{mbox-list-oflag } =\!/ \text{ "\NoUTF8" / "\UTF8Only"}
\]

\[
\text{return-option } =\!/ \text{ "UTF8"}
\]

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 which sends a message with UTF-8 headers to the server MUST include the UTF8 APPEND parameter. The ABNF for this APPEND parameter follows:

\[
\text{append-ext } =\!/ \text{ "UTF8"}
\]

A server which advertises UTF8=APPEND has to comply with the requirements of the IMAP base specification and RFC 2822 for message fetching. Mechanisms for 7-bit downgrading to help comply with the standards are discussed in Downgrading mechanism for Internationalized eMail Address (IMA) [I-D.yoneya-ima-downgrade].

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

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 which fails to comply with the formal syntax in RFC 3629 [RFC3629].

6. UTF8=ALL Capability

This 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 token.

7. UTF8=ONLY Capability

This 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 which find implementation of the UTF8 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 necessary to permit selection or examination of the mailbox in a backwards compatible fashion will become more difficult to support. Although it is hoped deployed IMAP servers do 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 base capability, the UTF8=ALL capability and the UTF8=APPEND capability. A server which 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. Minimal up-conversion is 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 encoded according to [TBD], address domains encoded according to IDNA [RFC3490], and MIME header encoding [RFC2047] of display-names and any RFC 2822 comments.


Up-conversion of MIME header encoding of the following headers MUST also be implemented: Subject, Date (RFC 2822 comments only), Comments, Keywords, Content-Description.

Server implementations also SHOULD up-convert all MIME body headers, SHOULD up-convert or remove the deprecated (and misused) name
parameter [RFC1341] on Content-Type and MUST up-convert the Content-Disposition filename parameter. 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 [RFC3501] base specification and RFC 2822 [RFC2822] 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 Internationalized eMail Address (IMA) [I-D.yoneya-ima-downgrade].

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", "UTF8=USER", "UTF8=APPEND", "UTF8=ALL", "UTF8=ONLY") to the IMAP4rev1 capability 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): none
   LIST-EXTENDED option description: Causes the LIST response to include mailboxes which mandate the UTF8 SELECT/EXAMINE parameter.
2. LIST-EXTENDED option name: UTF8ONLY

LIST-EXTENDED option type: SELECTION

Implied return options(s): none

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

Published specification: RFC XXXX, Section 3.4.1

Security considerations: RFC XXXX, Section 11

Intended usage: COMMON

Person an email address to contact for further information: see 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 XXXX, Section 3.4.1

Security considerations: RFC XXXX, Section 11

Intended usage: COMMON
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


[I-D.yeh-ima-utf8headers]
Yeh, J., "Internationalized Email Headers", draft-yeh-ima-utf8headers-01 (work in progress), February 2006.

[I-D.ietf-imapext-list-extensions]
Leiba, B. and A. Melnikov, "IMAP4 LIST Command Extensions", draft-ietf-imapext-list-extensions-17 (work in progress), April 2006.

12.2. Informative References


[I-D.newman-ima-pop]

[I-D.yoneya-ima-downgrade]
Yoneya, Y. and K. Fujiwara, "Downgrading mechanism for Internationalized eMail Address (IMA)", draft-yoneya-ima-downgrade-01 (work in progress), March 2006.
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 which 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 [I-D.newman-ima-pop].

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 author believes market forces, existing open-source software, and public conversion tables are sufficient to deploy the appropriate charsets.
Appendix B. Acknowledgments

TBD.

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
3401 Centrelake Dr., Suite 410
Ontario, CA  91761
US

Email: chris.newman@sun.com