POP3 Support for UTF-8

draft-ietf-eai-pop-00.txt

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 December 15, 2006.

Copyright Notice

Copyright (C) The Internet Society (2006).

Abstract

This specification extends the Post Office Protocol version 3 (POP3) to support un-encoded international characters in user names, mail addresses, message headers, and protocol-level textual error strings.
Table of Contents

1. Introduction ........................................... 3
   1.1. Conventions Used in this Document ................. 3
   1.2. Change History ..................................... 3
      1.2.1. Changes from draft-newman-ima-pop .............. 3
   1.3. Open Issues ........................................ 4
2. LANG Capability ....................................... 4
3. UTF8 Capability ........................................ 7
   3.1. USER Argument to UTF8 Capability ................... 7
   3.2. LST8 Argument to UTF8 Capability ................... 7
   3.3. TOP8 Argument to UTF8Capability ..................... 8
4. NO-RETR Capability ..................................... 8
5. Up-Conversion Server Requirements ....................... 9
6. Issues with UTF-8 Header Mail Drop ..................... 10
7. IANA Considerations ................................... 10
8. Security Considerations ................................ 10
9. References ............................................. 11
   9.1. Normative References ................................ 11
   9.2. Informative References ............................. 12
Appendix A. Design Rationale ............................. 12
Appendix B. Acknowledgments ............................... 14
Author’s Address ......................................... 15
Intellectual Property and Copyright Statements ............ 16
1. Introduction

This specification extends POP3 [RFC1939] using the POP3 Extension Mechanism [RFC2449] to permit un-encoded 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 login names outside the US-ASCII character set, and a mechanism to support UTF-8 protocol-level error strings in a language appropriate for the user.

Within this specification, the term up-conversion refers to converting a traditional 7-bit Internet message [RFC2822] with Message Header Extensions for Non-ASCII Text [RFC2047] and other 7-bit encodings to a message with UTF-8 headers [I-D.yeh-ima-utf8headers] and minimal use of 7-bit encodings. Down-conversion refers to the inverse process. One mechanism to perform down-conversion is described by Downgrading mechanism for Internationalized eMail Address [I-D.ietf-eai-downgrade].

1.1. 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 use the Augmented Backus-Naur Form (ABNF) [RFC4234] notation including the core rules defined in Appendix B of RFC 4234.

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.

1.2. Change History

This section describes the change history of this Internet draft and will be removed when/if this is published as an RFC.

1.2.1. Changes from draft-newman-ima-pop

- Change title to make this a WG document.
- Add LANG command and extension.
- Rename RET8 capability to UTF8 and add sub-sections for arguments.
o Add TOP8 command.

o Add definition of up-conversion and down-conversion.

o Some grammar fix-ups and section re-ordering based on RFC editor style.

1.3. Open Issues

The decision on how to handle UTF-8 in Received headers will impact the up-conversion requirements section.

2. LANG Capability

CAPA tag:

LANG

Arguments:

none

Added Commands:

LANG

Standard commands affected:

All

Announced states / possible differences:

both / no

Commands valid in states:

AUTHENTICATION, TRANSACTION

Specification reference:

this document

Discussion:

POP3 allows most +OK and -ERR server responses to include human-readable text that in some cases needs to be presented to the user. But that text is limited to US-ASCII by the POP3 specification [RFC1939]. The LANG capability and command permit a POP3 client to negotiate which language the server should use when sending human-readable text.

A server that advertises the LANG extension MUST use the language "i-default" as described in [RFC2277] as its default language until another supported language is negotiated by the client. A server
MUST include "i-default" as one of its supported languages.

The LANG command requests that human-readable text included in all subsequent +OK and -ERR responses be localized to a language matching the language range argument as described by section 2.5 of [RFC3066]. If the command succeeds, the server returns a +OK response followed by a single space, the exact RFC 3066 language tag selected, another space, and the rest of the line is human-readable text in the appropriate language. This and subsequent protocol-level human readable text is encoded in the UTF-8 charset.

If the command fails, the server returns a -ERR response and subsequent human-readable response text continues to use the language that was previously active (typically i-default).

The client MUST NOT use MUL (Multiple languages) or UND (Undetermined) language tags and the server MUST return -ERR if either tag is used. The special "*" language range argument indicates a request to use a language designated as preferred by the server administrator. The preferred language MAY vary based on the currently active user.

If no argument was given and the POP3 server issues a positive response, then the response given is multi-line. After the initial +OK, for each language tag the server supports, the POP3 server responds with a line for that language. This line is called a "language listing".

In order to simplify parsing, all POP3 servers are required to use a certain format for language listings. A language listing consists of the RFC 3066 language tag of the message, optionally followed by a single space and a human readable description of that language using the UTF-8 charset.
< The server defaults to using English i-default responses until the user explicitly changes the language. >

C: USER karen
S: +OK Hello, karen
C: PASS password
S: +OK karen’s maildrop contains 2 messages (320 octets)

< Client requested MUL language. Server MUST reply with -ERR >

C: LANG MUL
S: -ERR invalid language MUL

< A LANG command with no arguments is a request for a language listing. >

C: LANG
S: +OK Language listing follows:
S: en English
S: en-boont English Boontling dialect
S: de German
S: it Italian
S: i-default Default language
S: .

C: LANG
S: -ERR Server is unable to list languages

< Once the client changes the language, all responses will be in that language starting with the response to the LANG command. Note: the example does not include the correct character accents due to limitations of this document format. >

C: LANG fr
S: +OK fr La Language commande a ete execute avec success

< If a server does not support the requested primary language, responses will continue to be returned in the current language the server is using. >

C: LANG uga
S: -ERR Ce Language n’est pas supporte

C: LANG fr-ca
S: +OK fr La Language commande a ete execute avec success

C: LANG *
S: +OK fr La Language commande a ete execute avec success
3. UTF8 Capability

   CAPA tag:
     UTF8

   Arguments:
     USER, LST8, TOP8

   Added Commands:
     RET8, LST8, TOP8

   Standard commands affected:
     USER, PASS, APOP

   Announced states / possible differences:
     both / no

   Commands valid in states:
     TRANSACTION

   Specification reference:
     this document

   Discussion:

This capability adds UTF-8 content support to POP3. This capability always adds the "RET8" command to POP3. The RET8 command is identical to the RETR command, except that the retrieved message uses UTF-8 in headers [I-D.yeh-ima-utf8headers]. In addition, the 8bit content-transfer-encoding as defined in MIME section 2.8 [RFC2045] is explicitly permitted. The retrieved message MUST still be textual and otherwise formatted according to RFC 2822 [RFC2822] and MIME [RFC2045]. The MIME binary content-transfer-encoding is not permitted. Clients wishing to use binary MIME should implement IMAP4 [RFC3501] with the IMAP4 Binary Content Extension [RFC3516].

3.1. USER Argument to UTF8 Capability

   If the USER argument is included with this capability, that indicates the server accepts UTF-8 user names and passwords and applies SASLprep [RFC4013] to the arguments of the USER, PASS and APOP commands. A client that supports APOP and permits UTF-8 in user names or passwords MUST also implement SASLprep [RFC4013] on the user name and password used to compute the APOP digest.

3.2. LST8 Argument to UTF8 Capability

   If the LST8 argument is included with this capability, that indicates
the server implements the LST8 command. The LST8 command is identical to the LIST command except that the octet counts are the exact octet counts returned by the RET8 command. A POP3 client that uses RET8 MUST use LST8 instead of LIST if LST8 is advertised.

3.3. TOP8 Argument to UTF8 Capability

If the TOP8 argument is included with this capability, that indicates the server implements the TOP8 command. TOP8 is identical to TOP, except the headers are UTF-8.

4. NO-RETR Capability

CAPA tag:
   NO-RETR

Arguments:
   none

Added Commands:
   none

Standard commands affected:
   RETR, LIST, TOP

Announced states / possible differences:
   both / no

Commands valid in states:
   N/A

Specification reference:
   this document

Discussion:

This capability permits a POP3 server to advertise that it does not support the RETR, LIST or TOP commands. Any attempt to use any of these three commands results in an error response. As this is an incompatible change to POP3, a clear warning is necessary. POP3 clients that find implementation of the UTF8 capability problematic are encouraged to at least detect the NO-RETR capability and provide an informative error message to the end-user.

When a POP3 server runs on a UTF-8 header native mail drop, the downconversion step necessary to implement RETR in a backwards compatible fashion becomes more difficult to support. Although it is hoped
deployed POP3 servers do not advertise NO-RETR for some years, this capability is intended to minimize the disruption when legacy support finally goes away.

A server that advertises NO-RETR MUST advertise UTF8 with at least the LST8 argument and MUST NOT advertise TOP.

5. Up-Conversion Server Requirements

When a POP3 server uses a traditional mail drop that supports only 7-bit headers, it MUST support message header up-conversion for the RET8, LST8, and TOP8 commands. As POP3 clients are best when simple, the more up-conversion the server performs, the better. 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.

While this specification does not require it, server implementations are encouraged to up-convert all MIME body headers, and particularly the deprecated (and misused) name parameter [RFC1341] on Content-Type and the Content-Disposition [RFC2183] filename parameter. These may be encoded using the standard MIME parameter encoding [RFC2231] mechanism, or via non-standard use of MIME header encoding [RFC2047] in quoted strings.

Servers are also encouraged to up-convert the headers on embedded message/rfc822 body parts [TBD-ref]. Servers MAY convert the charset on MIME body parts to UTF-8, and MAY remove quoted-printable or base64 encodings as long as the resulting text complies with the requirements of the 8-bit content-transfer-encoding [RFC2045].
The POP3 server MUST NOT perform up-conversion of headers and content of multipart/signed [RFC1847], as well as Original-Recipient and Return-Path.

6. Issues with UTF-8 Header Mail Drop

When a POP3 server uses a mail drop that supports UTF-8 headers and it does not advertise the NO-RETR capability, it is the responsibility of the server to comply with the POP3 base specification [RFC1939] and RFC 2822 [RFC2822] with respect to the RETR, LIST, and TOP commands. Mechanisms for 7-bit downgrading to help comply with the standards are discussed in Downgrading mechanism for Internationalized eMail Address (IMA) [I-D.ietf-eai-downgrade].

A POP3 server with a mail drop that supports UTF-8 headers MUST comply with the RET8 protocol requirements implicit from Section 5. However, the code necessary for such compliance need not be part of the POP3 server itself in this case. For example, the minimal required up-conversion could be performed when a message is inserted into the POP3-accessible mail drop.

7. IANA Considerations

This adds three new capabilities ("UTF8", "LANG", and "NO-RETR") to the POP3 capability registry [RFC2449].

8. 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.

The "LANG *" command can reveal the existence and preferred language of a user to an active attacker probing the system if the active language changes in response to the USER, PASS, or APOP commands prior to validating the user’s credentials. Servers MUST implement a configuration to prevent this exposure.

It is possible for a man-in-the-middle attacker to insert a LANG command in the command stream thus making protocol-level diagnostic responses unintelligible to the user. A mechanism to integrity protect the session, such as TLS [RFC2595] can be used to defeat such attacks.
9. References

9.1. Normative References


9.2. Informative References


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 a parallel command set and permitting graceful migration of both client and server with minimal disruption is a deliberate choice. While a mechanism that makes RETR "just-send-UTF-8" might deploy faster, it would also create interoperability problems. The approach used prevents
interoperability problems until the NO-RETR mechanism is deployed. A
client command to cause a model switch could also work, but the
parallel command approach is cleaner given the small number of
commands.

The choice to make RET8 nearly identical to RETR is important to
minimize the code changes necessary in a client. An alternative
approach that permits binary MIME and uses a length-counted argument
would be architecturally superior but is dismissed due to the
migration problems it would cause. The IMAP4 Binary extension should
be sufficient for cases where binary MIME support is deemed
necessary.

LST8 is optional to minimize the cost of deploying UTF-8 support on a
legacy mail drop. The server load necessary to perform up-conversion
on every message in the mail drop to determine the LST8 octet-counts
would be prohibitively expensive when there’s no way to cache those
counts. The octet counts from the LIST command should be close
equal to the RET8 size for most POP3 user interfaces, and robust
POP3 clients already have to deal with LIST octet counts that don’t
match the actual size of the RETR result.

USER is optional because the implementation burden of SASLprep
[RFC4013] is not well understood and mandating such support in all
cases could negatively impact deployment.

The NO-RETR mechanism simplifies diagnosis of interoperability
problems when legacy support goes away. In the situation where
backwards compatibility is broken anyway, just-send-8 RETR has the
advantage that it might work with some legacy clients. However, the
difficulty of diagnosing interoperability problems caused by a just-
send-8 RETR mechanism is the reason the NO-RETR 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. While it would be desirable to require up-conversion of
attachment file names, the erroneous perception that MIME parsing is
difficult in combination with multiple deployed mechanisms for such
file names tip the balance.

Due to interoperability problems with RFC 2047 and limited deployment
of RFC 2231, it is hoped these 7-bit encoding mechanisms can be
deprecated in the future when UTF-8 header support becomes prevalent.
Aggressive conversion of these encodings to UTF-8 will help simplify
the infrastructure and improve interoperability in the future.

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. Specifically, use of an open-source charset
conversion library (such as ICU) is likely sufficient to fulfill this
recommendation.

While it is possible to provide useful examples for language
negotiation without support for non-ASCII characters, it is difficult
to provide useful examples for commands specifically designed to use
the UTF-8 charset un-encoded when the document format is limited to
US-ASCII. As a result, there are no plans to provide examples for
that part of the specification as long as this remains an
experimental proposal. However, implementers of this specification
are encouraged to provide examples to the document author for a
future revision.

This was deliberately written so the down-conversion specification is
not a normative reference. While this specification does reiterate
the requirements of the base POP3 specification with respect to
message format, no specific mechanism to achieve those requirements
is mandated.

Appendix B. Acknowledgments

Thanks to Randy Gellens, John Klensin, Tony Hansen and other EAI
working group participants who provided helpful suggestions and
interesting debate that improved this specification.
Author’s Address

Chris Newman
Sun Microsystems
3401 Centrelake Dr., Suite 410
Ontario, CA 91761
US

Email: chris.newman@sun.com