Overview and Framework for Internationalized Email
draft-ietf-eai-frmwrk-4952bis-08

Abstract

   Full use of electronic mail throughout the world requires that
   (subject to other constraints) people be able to use close variations
   on their own names (written correctly in their own languages and
   scripts) as mailbox names in email addresses. This document
   introduces a series of specifications that define mechanisms and
   protocol extensions needed to fully support internationalized email
   addresses. These changes include an SMTP extension and extension of
   email header syntax to accommodate UTF-8 data. The document set also
   includes discussion of key assumptions and issues in deploying fully
   internationalized email. This document is an update of RFC 4952; it
   reflects additional issues identified since that document was
   published.

Table of Contents

1. Introduction

   Note in Draft and to RFC Editor: The keyword represented in this
   document by "UTF8SMTPbis" (and in the XML source by &EAISMTPkeyword;)
   is a placeholder. The actual keyword will be assigned when the
   standards track SMTP extension in this series [RFC5336bis-SMTP] is
   approved for publication and should be substituted here. This
   paragraph should be treated as normative reference to that SMTP
   extension draft, creating a reference hold until it is approved by
   the IESG. The paragraph should be removed before RFC publication.

   In order to use internationalized email addresses, we need to
   internationalize both the domain part and the local part of email
   addresses. The domain part of email addresses is already
   internationalized [RFC5890], while the local part is not. Without
   the extensions specified in this document, the mailbox name is
   restricted to a subset of 7-bit ASCII [RFC5321]. Though MIME
   [RFC2045] enables the transport of non-ASCII data, it does not
   provide a mechanism for internationalized email addresses. In RFC
   2047 [RFC2047], MIME defines an encoding mechanism for some specific
   message header fields to accommodate non-ASCII data. However, it
   does not permit the use of email addresses that include non-ASCII
   characters. Without the extensions defined here, or some equivalent
   set, the only way to incorporate non-ASCII characters in any part of
   email addresses is to use RFC 2047 coding to embed them in what RFC
   5322 [RFC5322] calls the "display name" (known as a "name phrase" or
   by other terms elsewhere) of the relevant header fields. Information
   coded into the display name is invisible in the message envelope and,
   for many purposes, is not part of the address at all.

   This document is an update of RFC 4952 [RFC4952]; it reflects
   additional issues, shared terminology, and some architectural changes
   identified since that document was published.

   The pronouns "he" and "she" are used interchangeably to indicate a
   human of indeterminate gender.

   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 BCP 14, RFC 2119
   [RFC2119]. Although this document is Informational, those
   requirements are consistent with requirements specified in the
   Standards Track documents in this set as described in Section 5.

2. Role of This Specification

   This document presents the overview and framework for an approach to
   the next stage of email internationalization. This new stage

requires not only internationalization of addresses and header
fields, but also associated transport and delivery models. A prior
version of this specification, RFC 4952 [RFC4952], also provided an
introduction to a series of experimental protocols [RFC5335]
[RFC5336] [RFC5337] [RFC5504] [RFC5721] [RFC5738] [RFC5825]. This
revised form provides overview and conceptual information for the
standards-track successors of a subset of those protocols. Details
of the documents and the relationships among them appear in Section 5
and a discussion of what was learned from the Experimental protocols
and their implementations appears in Section 6.

Taken together, these specifications provide the details for a way to
implement and support internationalized email. The document itself
describes how the various elements of email internationalization fit
together and the relationships among the primary specifications
associated with message transport, header formats, and handling.

This document, and others that comprise the collection described
above, assume a reasonable familiarity with the basic Internet
electronic mail specifications and terminology [RFC5321][RFC5322] and
the MIME [RFC2045] and 8BITMIME [RFC1652] ones as well. While not
strictly required to implement this specification, a general
familiarity with the terminology and functions of IDNA
[RFC5890][RFC5891] [RFC5892][RFC5893] [RFC5894] are also assumed.

3. Problem Statement

Internationalizing Domain Names in Applications (IDNA) [RFC5890]
permits internationalized domain names, but deployment has not yet
reached most users. One of the reasons for this is that we do not
yet have fully internationalized naming schemes. Domain names are
just one of the various names and identifiers that are required to be
internationalized. In many contexts, until more of those identifiers
are internationalized, internationalized domain names alone have
little value.

Email addresses are prime examples of why it is not good enough to
just internationalize the domain name. As most observers have
learned from experience, users strongly prefer email addresses that
resemble names or initials to those involving seemingly meaningless
strings of letters or numbers. Unless the entire email address can
use familiar characters and formats, users will perceive email as
being culturally unfriendly. If the names and initials used in email
addresses can be expressed in the native languages and writing
systems of the users, the Internet will be perceived as more natural,
especially by those whose native language is not written in a subset
of a Roman-derived script.

Internationalization of email addresses is not merely a matter of
changing the SMTP envelope; or of modifying the From, To, and Cc
header fields; or of permitting upgraded Mail User Agents (MUAs) to
decode a special coding and respond by displaying local characters.
To be perceived as usable, the addresses must be internationalized
and handled consistently in all of the contexts in which they occur.
This requirement has far-reaching implications: collections of
patches and workarounds are not adequate. Even if they were
adequate, a workaround-based approach may result in an assortment of
implementations with different sets of patches and workarounds having
been applied with consequent user confusion about what is actually
usable and supported. Instead, we need to build a fully
internationalized email environment, focusing on permitting efficient
communication among those who share a language or other community.
That, in turn, implies changes to the mail header environment to
permit the full range of Unicode characters where that makes sense,
an SMTP Extension to permit UTF-8 [RFC3629] [RFC5198] mail addressing
and delivery of those extended header fields, support for
internationalization of delivery and service notifications [RFC3461]
[RFC3464], and (finally) a requirement for support of the 8BITMIME
SMTP Extension [RFC1652] so that all of these can be transported
through the mail system without having to overcome the limitation
that header fields do not have content-transfer-encodings.

4. Terminology

This document assumes a reasonable understanding of the protocols and
terminology of the core email standards as documented in [RFC5321]
and [RFC5322].

4.1. Mail User and Mail Transfer Agents

Much of the description in this document depends on the abstractions
of "Mail Transfer Agent" ("MTA") and "Mail User Agent" ("MUA").
However, it is important to understand that those terms and the
underlying concepts postdate the design of the Internet's email
architecture and the application of the "protocols on the wire"
principle to it. That email architecture, as it has evolved, and
that "on the wire" principle have prevented any strong and
standardized distinctions about how MTAs and MUAs interact on a given
origin or destination host (or even whether they are separate).

However, the term "final delivery MTA" is used in this document in a
fashion equivalent to the term "delivery system" or "final delivery
system" of RFC 5321. This is the SMTP server that controls the
format of the local parts of addresses and is permitted to inspect
and interpret them. It receives messages from the network for
delivery to mailboxes or for other local processing, including any

forwarding or aliasing that changes envelope addresses, rather than
relaying. From the perspective of the network, any local delivery
arrangements such as saving to a message store, handoff to specific
message delivery programs or agents, and mechanisms for retrieving
messages are all "behind" the final delivery MTA and hence are not
part of the SMTP transport or delivery process.

4.2. Address Character Sets

In this document, an address is "all-ASCII", or just an "ASCII
address", if every character in the address is in the ASCII character
repertoire [ASCII]; an address is "non-ASCII", or an "i18n-address",
if any character is not in the ASCII character repertoire. Such
addresses MAY be restricted in other ways, but those restrictions are
not relevant to this definition. The term "all-ASCII" is also
applied to other protocol elements when the distinction is important,
with "non-ASCII" or "internationalized" as its opposite.

The umbrella term to describe the email address internationalization
specified by this document and its companion documents is
"UTF8SMTPbis".
[[anchor3: Note in Draft: Keyword to be changed before publication.]]
For example, an address permitted by this specification is referred
to as a "UTF8SMTPbis (compliant) address".

Please note that, according to the definitions given here, the set of
all "all-ASCII" addresses and the set of all "non-ASCII" addresses
are mutually exclusive. The set of all addresses permitted when
UTF8SMTPbis appears is the union of these two sets.

4.3. User Types

An "ASCII user" (i) exclusively uses email addresses that contain
ASCII characters only, and (ii) cannot generate recipient addresses
that contain non-ASCII characters.

An "i18mail user" has one or more non-ASCII email addresses, or is
able to generate recipient addresses that contain non-ASCII
characters. Such a user may have ASCII addresses too; if the user
has more than one email account and a corresponding address, or more
than one alias for the same address, he or she has some method to
choose which address to use on outgoing email. Note that under this
definition, it is not possible to tell from an ASCII address if the
owner of that address is an i18mail user or not. (A non-ASCII
address implies a belief that the owner of that address is an i18mail
user.) There is no such thing as an "i18mail message"; the term
applies only to users and their agents and capabilities. In
particular, the use of non-ASCII message content is an integral part

of the MIME specifications [RFC2045] and does not require these
extensions (although it is compatible with them).

4.4. Messages

A "message" is sent from one user (sender) using a particular email
address to one or more other recipient email addresses (often
referred to just as "users" or "recipient users").

4.5. Mailing Lists

A "mailing list" is a mechanism whereby a message may be distributed
to multiple recipients by sending it to one recipient address. An
agent (typically not a human being) at that single address then
causes the message to be redistributed to the target recipients.
This agent sets the envelope return address of the redistributed
message to a different address from that of the original single
recipient message. Using a different envelope return address
(reverse-path) causes error (and other automatically generated)
messages to go to an error handling address.

Special provisions for managing mailing lists that might contain non-
ASCII addresses are discussed in a document that is specific to that
topic [EAI-Mailinglist] [RFCNNNNbis-MailingList].

4.6. Conventional Message and Internationalized Message

o A conventional message is one that does not use any extension
   defined in the SMTP extension document [RFC5336] or in the
   UTF8header specification [RFC5335], and is strictly conformant to
   RFC 5322 [RFC5322].

o An internationalized message is a message utilizing one or more of
   the extensions defined in this set of specifications, so that it
   is no longer conformant to the traditional specification of an
   email message or its transport.

4.7. Undeliverable Messages, Notification, and Delivery Receipts

As specified in RFC 5321, a message that is undeliverable for some
reason is expected to result in notification to the sender. This can
occur in either of two ways. One, typically called "Rejection",
occurs when an SMTP server returns a reply code indicating a fatal
error (a "5yz" code) or persistently returns a temporary failure
error (a "4yz" code). The other involves accepting the message
during SMTP processing and then generating a message to the sender,
typically known as a "Non-delivery Notification" or "NDN". Current
practice often favors rejection over NDNs because of the reduced

likelihood that the generation of NDNs will be used as a spamming
technique. The latter, NDN, case is unavoidable if an intermediate
MTA accepts a message that is then rejected by the next-hop server.

A sender MAY also explicitly request message receipts [RFC3461] that
raise the same issues for these internationalization extensions as
NDNs.

5. Overview of the Approach and Document Plan

This set of specifications changes both SMTP and the character
encoding of email message headers to permit non-ASCII characters to
be represented directly. Each important component of the work is
described in a separate document. The document set, whose members
are described below, also contains informational documents whose
purpose is to provide implementation suggestions and guidance for the
protocols.

In addition to this document, the following documents make up this
specification and provide advice and context for it.

o SMTP extensions. This document [RFC5336bis-SMTP] provides an SMTP
   extension (as provided for in RFC 5321) for internationalized
   addresses.

o Email message headers in UTF-8. This document [RFC5335bis-Hdrs]
   essentially updates RFC 5322 to permit some information in email
   message headers to be expressed directly by Unicode characters
   encoded in UTF-8 when the SMTP extension described above is used.
   This document, possibly with one or more supplemental ones, will
   also need to address the interactions with MIME, including
   relationships between UTF8SMTPbis and internal MIME headers and
   content types.

o Extensions to delivery status and notification handling to adapt
   to internationalized addresses [RFC5337bis-DSN].

o Extensions to the IMAP protocol to support internationalized
   message headers [RFC5738bis-IMAP].

o Parallel extensions to the POP protocol [RFC5721]
   [RFC5721bis-POP3].

6. Review of Experimental Results

The key difference between this set of protocols and the experimental
set that preceded them [RFC5335] [RFC5336] [RFC5337] [RFC5504]
[RFC5721] [RFC5738] [RFC5825] is that the earlier group provided a

mechanism for in-transit downgrading of messages (described in detail
in RFC 5504). That mechanism permitted, and essentially required,
that each non-ASCII address be accompanied by an all-ASCII
equivalent. That, in turn, raised security concerns associated with
pairing of addresses that could not be authenticated. It also
introduced the first incompatible change to Internet mail addressing
in many years, raising concerns about interoperability issues if the
new address forms "leaked" into legacy email implementations. The WG
concluded that the advantages of in-transit downgrading, were it
feasible operationally, would be significant enough to overcome those
concerns.

Operationally that turned out to not be the case, with
interoperability problems among initial implementations. Prior to
starting on the work that led to this set of specifications, the WG
concluded that the combination of requirements and long-term
implications of that earlier model were too complex to be
satisfactory and that work should move ahead without it.

7. Overview of Protocol Extensions and Changes

7.1. SMTP Extension for Internationalized Email Address

An SMTP extension, "UTF8SMTPbis" is specified as follows:

o Permits the use of UTF-8 strings in email addresses, both local
   parts and domain names.

o Permits the selective use of UTF-8 strings in email message
   headers (see Section 7.2).

o Requires that the server advertise the 8BITMIME extension
   [RFC1652] and that the client support 8-bit transmission so that
   header information can be transmitted without using a special
   content-transfer-encoding.

Some general principles affect the development decisions underlying
this work.

1. Email addresses enter subsystems (such as a user interface) that
    may perform charset conversions or other encoding changes. When
    the left hand side of the address includes characters outside the
    US-ASCII character repertoire, use of ASCII-compatible encoding
    (ACE) [RFC3492] [RFC5890] on the right hand side is discouraged
    to promote consistent processing of characters throughout the
    address.

2. An SMTP relay MUST

    *   Either recognize the format explicitly, agreeing to do so via
        an ESMTP option, or

    *   Reject the message or, if necessary, return a non-delivery
        notification message, so that the sender can make another
        plan.

3. If the message cannot be forwarded because the next-hop system
    cannot accept the extension it MUST be rejected or a non-delivery
    message MUST be generated and sent.

4. In the interest of interoperability, charsets other than UTF-8
    are prohibited in mail addresses and message headers being
    transmitted over the Internet. There is no practical way to
    identify multiple charsets properly with an extension similar to
    this without introducing great complexity.

Conformance to the group of standards specified here for email
transport and delivery requires implementation of the SMTP Extension
specification and the UTF-8 Header specification. If the system
implements IMAP or POP, it MUST conform to the i18n IMAP or POP
specifications respectively.

7.2. Transmission of Email Header Fields in UTF-8 Encoding

There are many places in MUAs or in a user presentation in which
email addresses or domain names appear. Examples include the
conventional From, To, or Cc header fields; Message-ID and
In-Reply-To header fields that normally contain domain names (but
that may be a special case); and in message bodies. Each of these
must be examined from an internationalization perspective. The user
will expect to see mailbox and domain names in local characters, and
to see them consistently. If non-obvious encodings, such as
protocol-specific ASCII-Compatible Encoding (ACE) variants, are used,
the user will inevitably, if only occasionally, see them rather than
"native" characters and will find that discomfiting or astonishing.
Similarly, if different codings are used for mail transport and
message bodies, the user is particularly likely to be surprised, if
only as a consequence of the long-established "things leak"
principle. The only practical way to avoid these sources of
discomfort, in both the medium and the longer term, is to have the
encodings used in transport be as similar to the encodings used in
message headers and message bodies as possible.

When email local parts are internationalized, they SHOULD be
accompanied by arrangements for the message headers to be in the

fully internationalized form. That form SHOULD use UTF-8 rather than
ASCII as the base character set for the contents of header fields
(protocol elements such as the header field names themselves are
unchanged and remain entirely in ASCII). For transition purposes and
compatibility with legacy systems, this can done by extending the
traditional MIME encoding models for non-ASCII characters in headers
[RFC2045] [RFC2231], but even these should be based on UTF-8. rather
than other encodings, if at all possible [IAB-idn-encoding].
However, the target is fully internationalized message headers, as
discussed in [RFC5335bis-Hdrs] and not an extended and painful
transition.

7.3. SMTP Service Extension for DSNs

The existing Draft Standard Delivery status notifications (DSNs)
specification [RFC3461] is limited to ASCII text in the machine
readable portions of the protocol. "International Delivery and
Disposition Notifications" [RFC5337bis-DSN] adds a new address type
for international email addresses so an original recipient address
with non-ASCII characters can be correctly preserved even after
downgrading. If an SMTP server advertises both the UTF8SMTPbis and
the DSN extension, that server MUST implement internationalized DSNs
including support for the ORCPT parameter specified in RFC 3461
[RFC3461].

8. Downgrading before and after SMTP Transactions

An important issue with these extensions is how to handle
interactions between systems that support non-ASCII addresses and
legacy systems that expect ASCII. There is, of course, no problem
with ASCII-only systems sending to those that can handle
internationalized forms because the ASCII forms are just a proper
subset. But, when systems that support these extensions send mail,
they MAY include non-ASCII addresses for senders, receivers, or both
and might also provide non-ASCII header information other than
addresses. If the extension is not supported by the first-hop system
(SMTP server accessed by the Submission server acting as an SMTP
client), message originating systems SHOULD be prepared to either
send conventional envelopes and message headers or to return the
message to the originating user so the message may be manually
downgraded to the traditional form, possibly using encoded words
[RFC2047] in the message headers. Of course, such transformations
imply that the originating user or system must have ASCII-only
addresses available for all senders and recipients. Mechanisms by
which such addresses may be found or identified are outside the scope
of these specifications as are decisions about the design of
originating systems such as whether any required transformations are
made by the user, the originating MUA, or the Submission server.

A somewhat more complex situation arises when the first-hop system
supports these extensions but some subsequent server in the SMTP
transmission chain does not. It is important to note that most cases
of that situation with forward-pointing addresses will be the result
of configuration errors: especially if it hosts non-ASCII addresses,
a final delivery MTA that accepts these extensions SHOULD NOT be
configured with lower-preference MX hosts that do not. When the only
non-ASCII address being transmitted is backward-pointing (e.g., in an
SMTP MAIL command), recipient configuration can not help in general.
On the other hand, alternate, all-ASCII, addresses for senders are
those most likely to be authoritatively known by the submission
environment or the sender herself. Consequently, if an intermediate
SMTP relay that requires these extensions then discovers that the
next system in the chain does not support them, it will have little
choice other than to reject or return the message.

As discussed above, downgrading to an ASCII-only form may occur
before or during the initial message submission. It might also occur
after the delivery to the final delivery MTA in order to accommodate
messages stores or IMAP or POP servers or clients that have different
capabilities than the delivery MTA. These two cases are discussed in
the subsections below.

8.1. Downgrading before or during Message Submission

The IETF has traditionally avoided specifying the precise behavior of
MUAs to provide maximum flexibility in the associated user
interfaces. The SMTP standard [RFC5321], Section 6.4, gives wide
latitude to MUAs and Submission servers as to what might be supplied
by the user as long as the result conforms with "on the wire"
standards once it is injected into the public Internet. In that
tradition, the discussion in the remainder of this section is
provided as general guidance rather than normative requirements.

It is likely that the most common cases in which a message that
requires these extensions is sent to a system that does not will
involve the combination of ASCII-only forward-pointing addresses with
a non-ASCII backward-pointing one. Until the extensions described
here have been universally implemented in the Internet email
environment, senders who prefer to use non-ASCII addresses (or raw
UTF-8 characters in header fields) even when their intended
recipients use and expect all-ASCII ones will need to be especially
careful about the error conditions that can arise, especially if they
are working in an environment in which non-delivery messages (or
other indications from submission servers) are routinely dropped or
ignored.

Perhaps obviously, the most convenient time to find an ASCII address

corresponding to an internationalized address is at the originating
MUA or closely-associated systems. This can occur either before the
message is sent or after the internationalized form of the message is
rejected. It is also the most convenient time to convert a message
from the internationalized form into conventional ASCII form or to
generate a non-delivery message to the sender if either is necessary.
At that point, the user has a full range of choices available,
including changing backward-pointing addresses, contacting the
intended recipient out of band for an alternate address, consulting
appropriate directories, arranging for translation of both addresses
and message content into a different language, and so on. While it
is natural to think of message downgrading as optimally being a
fully-automated process, we should not underestimate the capabilities
of a user of at least moderate intelligence who wishes to communicate
with another such user.

In this context, one can easily imagine modifications to message
submission servers (as described in RFC 4409 [RFC4409]) so that they
would perform downgrading operations or perhaps even upgrading ones.
Such operations would permit receiving messages with one or more of
the internationalization extensions discussed here and adapting the
outgoing message, as needed, to respond to the delivery or next-hop
environment the submission server encounters.

8.2. Downgrading or Other Processing After Final SMTP Delivery

When an email message is received by a final delivery MTA, it is
usually stored in some form. Then it is retrieved either by software
that reads the stored form directly or by client software via some
email retrieval mechanisms such as POP or IMAP.

The SMTP extension described in Section 7.1 provides protection only
in transport. It does not prevent MUAs and email retrieval
mechanisms that have not been upgraded to understand
internationalized addresses and UTF-8 message headers from accessing
stored internationalized emails.

Since the final delivery MTA (or, to be more specific, its
corresponding mail storage agent) cannot safely assume that agents
accessing email storage will always be capable of handling the
extensions proposed here, it MAY downgrade internationalized emails,
specially identify messages that utilize these extensions, or both.
If this is done, the final delivery MTA SHOULD include a mechanism to
preserve or recover the original internationalized forms without
information loss to support access by UTF8SMTPbis-aware agents.

9. Downgrading in Transit

   The base SMTP specification (Section 2.3.11 of RFC 5321 [RFC5321])
   states that "due to a long history of problems when intermediate
   hosts have attempted to optimize transport by modifying them, the
   local-part MUST be interpreted and assigned semantics only by the
   host specified in the domain part of the address". This is not a new
   requirement; equivalent statements appeared in specifications in 2001
   [RFC2821] and even in 1989 [RFC1123].

   Adherence to this rule means that a downgrade mechanism that
   transforms the local-part of an email address cannot be utilized in
   transit. It can only be applied at the endpoints, specifically by
   the MUA or submission server or by the final delivery MTA.

   One of the reasons for this rule has to do with legacy email systems
   that embed mail routing information in the local-part of the address
   field. Transforming the email address destroys such routing
   information. There is no way a server other than the final delivery
   server can know, for example, whether the local-part of
   user%foo@example.com is a route ("user" is reached via "foo") or
   simply a local address.

10. User Interface and Configuration Issues

   Internationalization of addresses and message headers, especially in
   combination with variations on character coding that are inherent to
   Unicode, may make careful choices of addresses and careful
   configuration of servers and DNS records even more important than
   they are for traditional Internet email. It is likely that, as
   experience develops with the use of these protocols, it will be
   desirable to produce one or more additional documents that offer
   guidance for configuration and interfaces. A document that discusses
   issues with mail user agents (MUAs), especially with regard to
   downgrading [EAI-MUA-issues], is expected to be developed in the EAI
   Working Group. The subsections below address some other issues.

10.1. Choices of Mailbox Names and Unicode Normalization

   It has long been the case that the email syntax permits choices about
   mailbox names that are unwise in practice if one actually intends the
   mailboxes to be accessible to a broad range of senders. The most-
   often-cited examples involve the use of case-sensitivity and tricky
   quoting of embedded characters in mailbox local parts. While these
   are permitted by the protocols and servers are expected to support
   them and there are special cases where they can provide value, taking
   advantage of those features is almost always bad practice unless the
   intent is to create some form of security by obscurity.

In the absence of these extensions, SMTP clients and servers are
constrained to using only those addresses permitted by RFC 5321. The
local parts of those addresses MAY be made up of any ASCII characters
except the control characters that 5321 prohibits, although some of
them MUST be quoted as specified there. It is notable in an
internationalization context that there is a long history on some
systems of using overstruck ASCII characters (a character, a
backspace, and another character) within a quoted string to
approximate non-ASCII characters. This form of internationalization
was permitted by RFC 821 [RFC0821] but is prohibited by RFC 5321
because it requires a backspace character (a prohibited C0 control).
Because RFC 5321 (and its predecessor, RFC 2821) prohibit the use of
this character in ASCII mailbox names and it is even more problematic
(for canonicalization and normalization reasons) in non-ASCII
strings, backspace MUST NOT appear in UTF8SMTPbis mailbox names.

For the particular case of EAI mailbox names, special attention MUST
be paid to Unicode normalization [Unicode-UAX15], in part because
Unicode strings may be normalized by other processes independent of
what a mail protocol specifies (this is exactly analogous to what may
happen with quoting and dequoting in traditional addresses).
Consequently, the following principles are offered as advice to those
who are selecting names for mailboxes:

o In general, it is wise to support addresses in Normalized form,
   using either Normalization Form NFC and, except in unusual
   circumstances, NFKC.

o It may be wise to support other forms of the same local-part
   string, either as aliases or by normalization of strings reaching
   the delivery server, in the event that the sender does not send
   the strings in normalized form.

o Stated differently and in more specific terms, the rules of the
   protocol for local-part strings essentially provide that:

   * Unnormalized strings are valid, but sufficiently bad practice
      that they may not work reliably on a global basis.

   * C0 (and presumably C1) controls (see The Unicode Standard
      [Unicode52]) are prohibited, the first in RFC 5321 and the
      second by an obvious extension from it [RFC5198].

   * Other kinds of punctuation, spaces, etc., are risky practice.
      Perhaps they will work, and SMTP receiver code is required to
      handle them, but creating dependencies on them in mailbox names
      that are chosen is usually a bad practice and may lead to
      interoperability problems.

11. Additional Issues

   This section identifies issues that are not covered, or not covered
   comprehensively, as part of this set of specifications, but that will
   require ongoing review as part of deployment of email address and
   header internationalization.

11.1. Impact on URIs and IRIs

   The mailto: schema [RFC2368] and discussed in the Internationalized
   Resource Identifier (IRI) specification [RFC3987] may need to be
   modified when this work is completed and standardized.

11.2. Use of Email Addresses as Identifiers

   There are a number of places in contemporary Internet usage in which
   email addresses are used as identifiers for individuals, including as
   identifiers to Web servers supporting some electronic commerce sites
   and in some X.509 certificates [RFC5280]. These documents do not
   address those uses, but it is reasonable to expect that some
   difficulties will be encountered when internationalized addresses are
   first used in those contexts, many of which cannot even handle the
   full range of addresses permitted today.

11.3. Encoded Words, Signed Messages, and Downgrading

   One particular characteristic of the email format is its persistency:
   MUAs are expected to handle messages that were originally sent
   decades ago and not just those delivered seconds ago. As such, MUAs
   and mail filtering software, such as that specified in Sieve
   [RFC5228], will need to continue to accept and decode header fields
   that use the "encoded word" mechanism [RFC2047] to accommodate non-
   ASCII characters in some header fields. While extensions to both
   POP3 [RFC1939] and IMAP [RFC3501] have been defined that include
   automatic upgrading of messages that carry non-ASCII information in
   encoded form -- including RFC 2047 decoding -- of messages by the
   POP3 [RFC5721bis-POP3] or IMAP [RFC5738bis-IMAP] server, there are
   message structures and MIME content-types for which that cannot be
   done or where the change would have unacceptable side effects.

   For example, message parts that are cryptographically signed, using
   e.g., S/MIME [RFC3851] or Pretty Good Privacy (PGP) [RFC3156], cannot
   be upgraded from the RFC 2047 form to normal UTF-8 characters without
   breaking the signature. Similarly, message parts that are encrypted
   may contain, when decrypted, header fields that use the RFC 2047
   encoding; such messages cannot be 'fully' upgraded without access to
   cryptographic keys.

Similar issues may arise if messages are signed and then subsequently
downgraded, e.g., as discussed in Section 8.1, and then an attempt is
made to upgrade them to the original form and then verify the
signatures. Even the very subtle changes that may result from
algorithms to downgrade and then upgrade again may be sufficient to
invalidate the signatures if they impact either the primary or MIME
bodypart headers. When signatures are present, downgrading MUST be
performed with extreme care if at all.

11.4. Other Uses of Local Parts

Local parts are sometimes used to construct domain labels, e.g., the
local part "user" in the address user@domain.example could be
converted into a vanity host user.domain.example with its Web space
at <http://user.domain.example> and the catchall addresses
any.thing.goes@user.domain.example.

Such schemes are obviously limited by, among other things, the SMTP
rules for domain names, and will not work without further
restrictions for other local parts such as the <utf8-local-part>
specified in [RFC5335bis-Hdrs]. Whether those limitations are
relevant to these specifications is an open question. It may be
simply another case of the considerable flexibility accorded to
delivery MTAs in determining the mailbox names they will accept and
how they are interpreted.

11.5. Non-Standard Encapsulation Formats

Some applications use formats similar to the application/mbox format
defined in [RFC4155] instead of the message/digest form described in
RFC 2046, Section 5.1.5 [RFC2046] to transfer multiple messages as
single units. Insofar as such applications assume that all stored
messages use the message/rfc822 format described in RFC 2046, Section
5.2.1 [RFC2046] with US-ASCII message headers, they are not ready for
the extensions specified in this series of documents and special
measures may be needed to properly detect and process them.

12. Key Changes From the Experimental Protocols and Framework

The original framework for internationalized email addresses and
headers was described in RFC 4952 and a subsequent set of
experimental protocol documents. Those relationships are described
in Section 3. The key architectural difference between the
experimental specifications and this newer set is that the earlier
specifications supported in-transit downgrading including providing
syntax and functions to support passing alternate, all-ASCII,
addresses with the non-ASCII ones and special headers to indicate the
downgraded status of messages. Those features were eliminated after

experimentation indicated that they were more complex and less
necessary than had been assumed earlier. Those issues are described
in more detail in Section 6 and Section 9.

13. IANA Considerations

This overview description and framework document does not contemplate
any IANA registrations or other actions. Some of the documents in
the group have their own IANA considerations sections and
requirements.

14. Security Considerations

Any expansion of permitted characters and encoding forms in email
addresses raises some risks. There have been discussions on so
called "IDN-spoofing" or "IDN homograph attacks". These attacks
allow an attacker (or "phisher") to spoof the domain or URLs of
businesses. The same kind of attack is also possible on the local
part of internationalized email addresses. It should be noted that
the proposed fix involving forcing all displayed elements into
normalized lower-case works for domain names in URLs, but not email
local parts since those are case sensitive.

Since email addresses are often transcribed from business cards and
notes on paper, they are subject to problems arising from confusable
characters (see [RFC4690]). These problems are somewhat reduced if
the domain associated with the mailbox is unambiguous and supports a
relatively small number of mailboxes whose names follow local system
conventions. They are increased with very large mail systems in
which users can freely select their own addresses.

The internationalization of email addresses and message headers must
not leave the Internet less secure than it is without the required
extensions. The requirements and mechanisms documented in this set
of specifications do not, in general, raise any new security issues.

They do require a review of issues associated with confusable
characters -- a topic that is being explored thoroughly elsewhere
(see, e.g., RFC 4690 [RFC4690]) -- and, potentially, some issues with
UTF-8 normalization, discussed in RFC 3629 [RFC3629], and other
transformations. Normalization and other issues associated with
transformations and standard forms are also part of the subject of
work described elsewhere [RFC5198] [RFC5893] [IAB-idn-encoding].

Some issues specifically related to internationalized addresses and
message headers are discussed in more detail in the other documents
in this set. However, in particular, caution should be taken that
any "downgrading" mechanism, or use of downgraded addresses, does not

inappropriately assume authenticated bindings between the
internationalized and ASCII addresses. Expecting and most or all
such transformations prior to final delivery be done by systems that
are presumed to be under the administrative control of the sending
user ameliorates the potential problem somewhat as compared to what
it would be if the relationships were changed in transit.

The new UTF-8 header and message formats might also raise, or
aggravate, another known issue. If the model creates new forms of an
'invalid' or 'malformed' message, then a new email attack is created:
in an effort to be robust, some or most agents will accept such
message and interpret them as if they were well-formed. If a filter
interprets such a message differently than the MUA used by the
recipient, then it may be possible to create a message that appears
acceptable under the filter's interpretation but that should be
rejected under the interpretation given to it by that MUA. Such
attacks already exist for existing messages and encoding layers,
e.g., invalid MIME syntax, invalid HTML markup, and invalid coding of
particular image types.

In addition, email addresses are used in many contexts other than
sending mail, such as for identifiers under various circumstances
(see Section 11.2). Each of those contexts will need to be
evaluated, in turn, to determine whether the use of non-ASCII forms
is appropriate and what particular issues they raise.

This work will clearly affect any systems or mechanisms that are
dependent on digital signatures or similar integrity protection for
email message headers (see also the discussion in Section 11.3).
Many conventional uses of PGP and S/MIME are not affected since they
are used to sign body parts but not message headers. On the other
hand, the developing work on domain keys identified mail (DKIM)
[RFC5863] will eventually need to consider this work and vice versa:
while this specification does not address or solve the issues raised
by DKIM and other signed header mechanisms, the issues will have to
be coordinated and resolved eventually if the two sets of protocols
are to co-exist. In addition, to the degree to which email addresses
appear in PKI (Public Key Infrastructure) certificates, standards
addressing such certificates will need to be upgraded to address
these internationalized addresses. Those upgrades will need to
address questions of spoofing by look-alikes of the addresses
themselves.

15. Acknowledgments

This document is an update to, and derived from, RFC 4952. This
document would have been impossible without the work and
contributions acknowledged in it. The present document benefited

significantly from discussions in the EAI WG and elsewhere after RFC
4952 was published, especially discussions about the experimental
versions of other documents in the internationalized email
collection, and from RFC errata on RFC 4952 itself.

Special thanks are due to Ernie Dainow for careful reviews and
suggested text in this version.

16. References

16.1. Normative References

[ASCII]   American National Standards Institute
                         (formerly United States of America
                         Standards Institute), "USA Code for
                         Information Interchange", ANSI X3.4-1968,
                         1968.

                         ANSI X3.4-1968 has been replaced by newer
                         versions with slight modifications, but the
                         1968 version remains definitive for the
                         Internet.

[RFC1652]   Klensin, J., Freed, N., Rose, M.,
                         Stefferud, E., and D. Crocker, "SMTP
                         Service Extension for 8bit-MIMEtransport",
                         RFC 1652, July 1994.

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

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

[RFC5321]   Klensin, J., "Simple Mail Transfer
                         Protocol", RFC 5321, October 2008.

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

[RFC5335bis-Hdrs]   Yang, A. and S. Steele, "Internationalized
                         Email Headers", July 2010, <https://
                         datatracker.ietf.org/doc/
                         draft-ietf-eai-rfc5335bis/>.

[RFC5336bis-SMTP]   Yao, J. and W. Mao, "SMTP Extension for

                                   Internationalized Email Address",
                                   August 2010, <https://datatracker.ietf.org/
                                   doc/draft-ietf-eai-rfc5336bis/>.

   [RFC5337bis-DSN]   Not yet posted?, "Internationalized
                                   Delivery Status and Disposition
                                   Notifications", Unwritten waiting for I-D,
                                   2010.

   [RFC5721bis-POP3]   Not yet posted?, "POP3 Support for UTF-8",
                                   Unwritten waiting for I-D, 2010.

   [RFC5738bis-IMAP]   Not yet posted?, "IMAP Support for UTF-8",
                                   Unwritten waiting for I-D, 2010.

   [RFC5890]   Klensin, J., "Internationalized Domain
                                   Names for Applications (IDNA): Definitions
                                   and Document Framework", RFC 5890,
                                   August 2010.

   [RFCNNNNbis-MailingList]   Not yet posted?, "Mailing Lists and
                                   Internationalized Email Addresses", First
                                   Version still not in RFC Editor queue https
                                   ://datatracker.ietf.org/doc/
                                   draft-ietf-eai-mailinglist/,
                                   Unwritten waiting for I-D, 2010.

16.2. Informative References

   [EAI-MUA-issues]   EAI WG, "Still-unnamed proposed document on
                                   MUA issues", Not assigned or agreed to yet,
                                   2011.

                                   Note to IESG and RFC Editor: While there is
                                   provision for a document on this subject in
                                   the WG Charter, there is, as yet, no plan
                                   for producing it or even for adding it to
                                   the WG's task list with benchmarks. If the
                                   present document is approved for
                                   publication before the is at least a title
                                   and author(s) for an I-D, the citation and
                                   reference should simply be dropped.

   [EAI-Mailinglist]   Gellens, R., "Mailing Lists and
                                   Internationalized Email Addresses",
                                   June 2010, <https://datatracker.ietf.org/
                                   doc/draft-ietf-eai-mailinglist/>.

[IAB-idn-encoding]   Thaler, D., Klensin, J., and S. Cheshire,
                          "IAB Thoughts on Encodings for
                          Internationalized Domain Names", 2010,.

[RFC0821]   Postel, J., "Simple Mail Transfer
                          Protocol", STD 10, RFC 821, August 1982.

[RFC1123]   Braden, R., "Requirements for Internet
                          Hosts - Application and Support", STD 3,
                          RFC 1123, October 1989.

[RFC1939]   Myers, J. and M. Rose, "Post Office
                          Protocol - Version 3", STD 53, RFC 1939,
                          May 1996.

[RFC2033]   Myers, J., "Local Mail Transfer Protocol",
                          RFC 2033, October 1996.

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

[RFC2046]   Freed, N. and N. Borenstein, "Multipurpose
                          Internet Mail Extensions (MIME) Part Two:
                          Media Types", RFC 2046, November 1996.

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

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

[RFC2368]   Hoffman, P., Masinter, L., and J. Zawinski,
                          "The mailto URL scheme", RFC 2368,
                          July 1998.

[RFC2821]   Klensin, J., "Simple Mail Transfer
                          Protocol", RFC 2821, April 2001.

[RFC3156]   Elkins, M., Del Torto, D., Levien, R., and
                          T. Roessler, "MIME Security with OpenPGP",

                              RFC 3156, August 2001.

   [RFC3461]   Moore, K., "Simple Mail Transfer Protocol
                              (SMTP) Service Extension for Delivery
                              Status Notifications (DSNs)", RFC 3461,
                              January 2003.

   [RFC3464]   Moore, K. and G. Vaudreuil, "An Extensible
                              Message Format for Delivery Status
                              Notifications", RFC 3464, January 2003.

   [RFC3492]   Costello, A., "Punycode: A Bootstring
                              encoding of Unicode for Internationalized
                              Domain Names in Applications (IDNA)",
                              RFC 3492, March 2003.

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

   [RFC3851]   Ramsdell, B., "Secure/Multipurpose Internet
                              Mail Extensions (S/MIME) Version 3.1
                              Message Specification", RFC 3851,
                              July 2004.

   [RFC3987]   Duerst, M. and M. Suignard,
                              "Internationalized Resource Identifiers
                              (IRIs)", RFC 3987, January 2005.

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

   [RFC4409]   Gellens, R. and J. Klensin, "Message
                              Submission for Mail", RFC 4409, April 2006.

   [RFC4690]   Klensin, J., Faltstrom, P., Karp, C., and
                              IAB, "Review and Recommendations for
                              Internationalized Domain Names (IDNs)",
                              RFC 4690, September 2006.

   [RFC4952]   Klensin, J. and Y. Ko, "Overview and
                              Framework for Internationalized Email",
                              RFC 4952, July 2007.

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

   [RFC5228]   Guenther, P. and T. Showalter, "Sieve: An
                          Email Filtering Language", RFC 5228,
                          January 2008.

   [RFC5280]   Cooper, D., Santesson, S., Farrell, S.,
                          Boeyen, S., Housley, R., and W. Polk,
                          "Internet X.509 Public Key Infrastructure
                          Certificate and Certificate Revocation List
                          (CRL) Profile", RFC 5280, May 2008.

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

   [RFC5336]   Yao, J. and W. Mao, "SMTP Extension for
                          Internationalized Email Addresses",
                          RFC 5336, September 2008.

   [RFC5337]   Newman, C. and A. Melnikov,
                          "Internationalized Delivery Status and
                          Disposition Notifications", RFC 5337,
                          September 2008.

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

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

   [RFC5738]   Resnick, P. and C. Newman, "IMAP Support
                          for UTF-8", RFC 5738, March 2010.

   [RFC5825]   Fujiwara, K. and B. Leiba, "Displaying
                          Downgraded Messages for Email Address
                          Internationalization", RFC 5825,
                          April 2010.

   [RFC5863]   Hansen, T., Siegel, E., Hallam-Baker, P.,
                          and D. Crocker, "DomainKeys Identified Mail
                          (DKIM) Development, Deployment, and
                          Operations", RFC 5863, May 2010.

   [RFC5891]   Klensin, J., "Internationalized Domain
                          Names in Applications (IDNA): Protocol",
                          RFC 5891, August 2010.

   [RFC5892]   Faltstrom, P., "The Unicode Code Points and

                              Internationalized Domain Names for
                              Applications (IDNA)", RFC 5892,
                              August 2010.

   [RFC5893]   Alvestrand, H. and C. Karp, "Right-to-Left
                              Scripts for Internationalized Domain Names
                              for Applications (IDNA)", RFC 5893,
                              August 2010.

   [RFC5894]   Klensin, J., "Internationalized Domain
                              Names for Applications (IDNA): Background,
                              Explanation, and Rationale", RFC 5894,
                              August 2010.

   [Unicode-UAX15]   The Unicode Consortium, "Unicode Standard
                              Annex #15: Unicode Normalization Forms",
                              March 2008,
                              <http://www.unicode.org/reports/tr15/>.

   [Unicode52]   The Unicode Consortium. The Unicode
                              Standard, Version 5.2.0, defined by:, "The
                              Unicode Standard, Version 5.2.0", (Mountain
                              View, CA: The Unicode Consortium,
                              2009. ISBN 978-1-936213-00-9)., <http://
                              www.unicode.org/versions/Unicode5.2.0/>.

Appendix A. Change Log

   [[RFC Editor: Please remove this section prior to publication.]]

A.1. Changes between -00 and -01

   o Because there has been no feedback on the mailing list, updated
      the various questions to refer to this version as well.

   o Reflected RFC Editor erratum #1507 by correcting terminology for
      headers and header fields and distinguishing between "message
      headers" and different sorts of headers (e.g., the MIME ones).

A.2. Changes between -01 and -02

   Note that section numbers in the list that follows may refer to -01
   and not -02.

   o Discussion of RFC 5825 ("downgraded display") has been removed per
      the earlier note and on-list discussion. Any needed discussion
      about reconstructed messages will need to appear in the IMAP and
      POP documents. However, the introductory material has been

reworded to permit keeping 5504 and 5825 on the list there, without which the back chain would not be complete. For consistency with this change, 5504 and 5825 have been added to the "Obsoletes" list (as far as I know, an Informational spec can obsolete or update Experimental ones, so no downref problem here --JcK).

o Reference to alternate addresses dropped from (former) Section 7.1.

o Reference to RFC 5504 added to (former) Section 8 for completeness.

o Ernie's draft comments added (with some minor edits) to replace the placeholder in (former) Section 9 ("Downgrading in Transit"). It is intended to capture at least an introduction the earlier discussions of algorithmic downgrading generally and ACE/Punycode transformations in particular. Anyone who is unhappy with it should say so and propose alternate text. RSN.

o In the interest of clarity and consistency with the terminology in Section 4.1, all uses of "final delivery SMTP server" and "final delivery server" have been changed to "final delivery MTA".

o Placeholder at the end of Section 2 has been removed and the text revised to promise less. The "Document Plan" (Section 5) has been revised accordingly. We need to discuss this at IETF 78 if not sooner.

o Sections 5 and 6 have been collapsed into one -- there wasn't enough left in the former Section 5 to justify a separate section.

o Former Section 11.1 has been dropped and the DSN document moved up into the "Document Plan" as suggested earlier.

o Section 12, "Experimental Targets", has been removed.

o Updated references for the new version EAI documents and added placeholders for all of the known remaining drafts that will become part of the core EAI series but that have not been written.

o Inserted an additional clarification about the relationship of these extensions to non-ASCII messages.

o Changed some normative/informative reference classifications based on review of the new text.

   o Removed references to the pre-EAI documents that were cited for
      historical context in 4952.

   o Got rid of a remaining pointer to address downgrading in the
      discussion of an updated MAILTO URI.

   o Minor additional editorial cleanups and tuning.

A.3. Changes between -02 and -03

   o Inserted paragraph clarifying the status of the UTF8SMTPbis
      keyword as a result of discussion prior to and during IETF 79.

   o Adjusted some references including adding an explicit citation of
      RFC 821.

   o Removed the discussion of the experimental work from an inline
      aside to a separate section, Section 6.

   o Rewrote the discussion of configuration errors in MX setups to
      make it clear that they are an issue with forward-pointing
      addresses only and improved the discussion of backward-pointing
      addresses.

   o Removed some now-obsolete placeholder notes and resolved the
      remaining one to a dangling reference.

A.4. Changes between -03 and -04

   o Several minor editorial changes.

   o Added a discussion of the relationship to the base email, MIME,
      and IDNA specifications.

A.5. Changes between -04 and -05

   o Several more minor editorial changes.

A.6. Changes between -05 and -06

   o Corrections to more precisely reflect RFC 2119 language
      requirements and closely-related issues..

A.7. Changes between -06 and -07

   o Added a new section (now Section 12) to explicitly discuss the
      changes from the previous version.

o Removed the discussion of LMTP from Section 11; it is more
   appropriately placed in the SMTP Extension document (5336bis).

A.8. Changes between -07 and -08 (after IETF Last Call)

o Modified Section 7.2 to make the last paragraph less tentative and
   more clear.

o Modified Section 8.1 to add an introductory paragraph that
   clarifies what the IETF does and does not specify about email
   protocols.

Authors' Addresses

John C KLENSIN
1770 Massachusetts Ave, #322
Cambridge, MA 02140
USA

Phone: +1 617 491 5735
EMail: john-ietf@jck.com

YangWoo KO
ICU
119 Munjiro
Yuseong-gu, Daejeon 305-732
Republic of Korea

EMail: yw@mrko.pe.kr