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 that reflects additional issues identified since that document was published.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on January 4, 2011.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.
Table of Contents

1. Introduction ................................................. 4
2. Role of This Specification .................................. 4
3. Problem Statement ............................................ 5
4. Terminology .................................................. 6
   4.1. Mail User and Mail Transfer Agents .................. 6
   4.2. Address Character Sets ................................ 7
   4.3. User Types ............................................. 7
   4.4. Messages ............................................... 7
   4.5. Mailing Lists ......................................... 8
   4.6. Undeliverable Messages and Notification ............. 8
5. Overview of the Approach .................................... 8
6. Document Plan ............................................... 9
7. Overview of Protocol Extensions and Changes ............... 9
   7.1. SMTP Extension for Internationalized Email Address ... 9
   7.2. Transmission of Email Header Fields in UTF-8 Encoding .. 10
8. Downgrading before and after SMTP Transactions ............ 11
   8.1. Downgrading before or during Message Submission ...... 12
   8.2. Downgrading or Other Processing After Final SMTP Delivery ............................................. 13
9. Downgrading in Transit ....................................... 13
10. User Interface and Configuration Issues ................... 13
   10.1. Choices of Mailbox Names and Unicode Normalization .... 14
11. Additional Issues .......................................... 15
   11.1. Impact on URIs and IRIs ............................... 15
   11.2. Interaction with Delivery Notifications ............. 15
   11.3. Use of Email Addresses as Identifiers ............... 16
   11.4. Encoded Words, Signed Messages, and Downgrading ...... 16
   11.5. LMTP .................................................. 16
   11.6. Other Uses of Local Parts ............................. 16
   11.7. Non-Standard Encapsulation Formats .................. 17
12. Experimental Targets ....................................... 17
13. IANA Considerations ....................................... 17
14. Security Considerations ................................... 17
15. Acknowledgements ......................................... 19
16. References ............................................... 19
   16.1. Normative References ............................... 19
   16.2. Informative References ............................. 20
Appendix A. Change Log ....................................... 22
   A.1. Changes between -00 and -01 .......................... 23
1. Introduction

[[anchor1: Note to EAI WG: these two initial drafts are intended to initiate discussion on what should, and should not, be in the Framework document and how we want those topics covered. As such, it is more of an intermediate draft between RFC 4952 and the first draft of 4952bis that could be a Last Call candidate. If we are going to keep the rather aggressive schedule we agreed to in the charter, we need to have enough discussion on critical-path points that a revision suitable (at least) for final review prior to Last Call can be posted before the 12 July I-D cutoff. For that to happen, we should have enough discussion to start determining consensus within the next ten days. So, focused comments and soon, please.]]

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] that 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", "SHALL", "REQUIRED", "SHOULD", "RECOMMENDED", and "MAY" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 document provides an overview and conceptual information for the standards-track successors of those protocols. Details of the documents and the relationships among them appear 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 various documents are involved.

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 of us 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] mail addressing and delivery of those extended header fields, 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 "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".  

[[anchor7: 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. 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.

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").

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].

An internationalized message is a message utilizing one or more of the extensions defined in this specification or in the UTF8header specification [RFC5335], so that it is no longer conformant to the RFC 5322 specification of a message.

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].

4.6. Undeliverable Messages and Notification

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.

5. Overview of the Approach

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 in the next section, also contains informational documents whose purpose is to provide implementation suggestions and guidance for the protocols.
6. Document Plan

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

[[anchor12: ... Note to WG: if we actually include a list here, the result will be that this document can be approved, but not published, until those documents on the list are complete. I’m inclined to list the SMTP extension and headers documents only and hand-wave about the rest, but we need to discuss. Versions -00 and -01 simply refer to the current Experimental documents --Editor.]]

- SMTP extensions. This document [RFC5336] provides an SMTP extension (as provided for in RFC 5321) for internationalized addresses.

- Email message headers in UTF-8. This document [RFC5335] 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.

- Extensions to the IMAP protocol to support internationalized message headers [RFC5738].

- Parallel extensions to the POP protocol [RFC5721].

- Description of internationalization changes for delivery notifications (DSNs) [RFC5337].

7. Overview of Protocol Extensions and Changes

7.1. SMTP Extension for Internationalized Email Address

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

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

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

- 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 punycode 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 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, including recognition of the keywords associated with alternate addresses, 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, it seems clear that 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 will remain entirely in ASCII). For transition purposes and compatibility with legacy systems, this can be done by extending the encoding models of [RFC2045] and [RFC2231]. However, the target is fully internationalized message headers, as discussed in [RFC5335] and not an extended and painful transition.

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 will be the result of configuration errors: especially if it hosts non-ASCII addresses, a final delivery server that accepts these extensions should not be configured with lower-preference MX hosts that do not. While the experiments that preceded these specifications included a mechanism for passing backup ASCII addresses to intermediate relay systems and having those systems alter the relevant message header fields and substitute the addresses, the requirements and long-term implications of that system proved too complex to be satisfactory. Consequently, if an intermediate SMTP relay that is transmitting a message that requires these extensions and 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

Perhaps obviously, the most convenient time to find an ASCII address corresponding to an internationalized address is at the originating MUA. 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 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 [RFC4409]) so that they would perform downgrading, or perhaps even upgrading, operations, 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 it encounters.
8.2. Downgrading or Other Processing After Final SMTP Delivery

When an email message is received by a final delivery SMTP server, 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 SMTP server (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 either downgrade internationalized emails or specially identify messages that utilize these extensions, or both. If this is done, the final delivery SMTP server 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

[[anchor16: Note in Draft and Question for the WG: We could discuss the various issues with in-transit downgrading including the complexities of carrying backup addresses, the problems that motivated the "don’t mess with addresses in transit" (paraphrased, obviously) rule in RFC 5321 and friends, and so on. Or we could omit it (and this section). Pragmatically, I think it would take us some time to reach consensus on what, exactly, should be said and that might delay progress. But input is clearly needed -- if it is not received before we prepared -02, this section will simply be dropped.]]

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, 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.

In the absence of this extension, 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 but is prohibited by RFC 5321 because it requires a backspace character (a prohibited C0 control). The practice SHOULD be phased out as this extension becomes widely deployed but backward-compatibility considerations may require that it continue to be recognized.

For the particular case of EAI mailbox names, special attention must be paid to Unicode normalization, 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:

- In general, it is wise for servers to provide addresses only in Normalized form and to normalize strings on receipt, using either Normalization Form NFC and, except in unusual circumstances, NFKC.
  
  [[anchor19: Note in Draft: "Normalize on receipt" is consistent with the recommendations in draft-iab-i18n-encoding. The issue with NFKC is that some of the characters mapped out may be significant, especially in personal names. Anyone with objections should speak up. Soon.]]

- 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.

- 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) are
    prohibited, the first in RFC 5321 and the second by an obvious
    extension from it.

  * 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 defined in [RFC2368] and discussed in the
Internationalized Resource Identifier (IRI) specification [RFC3987]
may need to be modified when this work is completed and standardized.
In particular, providing an alternate address as part of a mailto:
URI may require some fairly careful work on the syntax of that URI.

11.2. Interaction with Delivery Notifications

The advent of UTF8SMTPbis will make necessary consideration of the
interaction with delivery notification mechanisms, including the
ASCII-only SMTP extension for requesting delivery notifications
(DSNs) [RFC3461], and the format of delivery notifications [RFC3464].
A new document, "International Delivery and Disposition
Notifications" [RFC5337] 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.
11.3. 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. 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.

11.4. 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 and IMAP have been proposed to enable automatic EAI-upgrade -- including RFC 2047 decoding -- of messages by the POP3 or 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.

11.5. LMTP

LMTP [RFC2033] may be used as the final delivery agent. In such cases, LMTP may be arranged to deliver the mail to the mail store. The mail store may not have UTF8SMTPbis capability. LMTP need to be updated to deal with these situations.

11.6. 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 [RFC5335]. Whether this issue is 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.

11.7. Non-Standard Encapsulation Formats

Some applications use formats similar to the application/mbox format defined in [RFC4155] instead of the message/digest RFC 2046, Section 5.1.5 [RFC2046] form to transfer multiple messages as single units. Insofar as such applications assume that all stored messages use the message/rfc822 RFC 2046, Section 5.2.1 [RFC2046] format 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. Experimental Targets

[[anchor26: Note in draft: this section is left in this draft for convenience in review. It will be removed with -02.]]

In addition to the simple question of whether the model outlined here can be made to work in a satisfactory way for upgraded systems and provide adequate protection for un-upgraded ones, we expect that actually working with the systems will provide answers to two additional questions: what restrictions such as character lists or normalization should be placed, if any, on the characters that are permitted to be used in address local-parts and how useful, in practice, will downgrading turn out to be given whatever restrictions and constraints that must be placed upon it.

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., [RFC4690]) -- and, potentially, some issues with UTF-8
normalization, discussed in [RFC3629], and other transformations.
Normalization and other issues associated with transformations and
standard forms are also part of the subject of ongoing work discussed
in [RFC5198], in [RFC5893] and elsewhere.

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 final MUA, then it may
be possible to create a message that appears acceptable under the
filter’s interpretation but should be rejected under the
interpretation given to it by the final 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 11.3). 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 11.4). 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. Acknowledgements

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.

16. References

16.1. Normative References


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


[JET-IMA]         Yao, J. and J. Yeh, "Internationalized eMail Address (IMA)", Work in Progress, June 2005.


[RFC2047]         Moore, K., "MIME (Multipurpose Internet Mail

Klensin & Ko    Expires January 4, 2011   [Page 20]
Extensions) Part Three: Message Header


[RFC5198] Klensin, J. and M. Padlipsky, "Unicode Format


Appendix A. Change Log

[This section is for internal use only. Please remove this section prior to publication.]
A.1. Changes between -00 and -01

- Because there has been no feedback on the mailing list, updated the various questions to refer to this version as well.
- 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).
- Merged earlier sections 4.4 and 4.6 into an expanded Section 4.4.
- Merged earlier Section 11.6 into Section 11.2 and eliminated the note in draft.
- Eliminated former last paragraph of Section 11.4 as an artifact of in-transit downgrading.
- Updated a few references.

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