Overview and Framework for Internationalized Email
draft-klensin-ima-framework-01.txt

Status of this Memo

Copyright Notice

Abstract

   Full use of electronic mail throughout the world requires that people
   be able to use their own names, written correctly in their own
   languages and scripts, as mailbox names in email addresses. This
   document introduces a series of specifications and operational
   suggestions 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 will include
discussion of key assumptions and issues in deploying fully
internationalized email.

Table of Contents

1. Introduction

   In order to use internationalized email addresses, we need to
   internationalize both domain part and local part of email address.
   The domain part of email addresses is already internationalized
   [RFC3490], while the local part is not. Without these extensions,
   the mailbox name is restricted to a subset of 7-bit ASCII in
   [RFC2821]. Though MIME enables the transport of non-ASCII data, it
   does not provide a mechanism for internationalized email address.
   [RFC2047] defines an encoding mechanism for some specific message
   header fields to accommodate non-ASCII data. However, it does not
   address the issue 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 email
   addresses is to use RFC2047 coding to embed them in the "name
   phrases" of the relevant headers. Of course, that type of coding is
   invisible in the message envelope and would not be considered by many
   to be part of the address at all.

1.1. 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 headers, but
   also associated transport and delivery models. The history of
   developments and design ideas leading to this specification is
   described in [I18Nemail-history].

   This document describes how the various elements of email
   internationalization fit together and provides a roadmap for
   navigating the various documents involved.

1.2. Problem statement

   [[anchor1: Note in draft: this section needs very significant
   reworking for both content and presentation. Changed with -01c, but
   may still not be good enough]]

   Though domain names are already internationalized, the
   internationalized forms are far from general adoption by ordinary
   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.

   Email addresses are a particularly important example of where
   internationalization of domain names alone is not sufficient. Unless
   email addresses are presented to the user in familiar characters and

formats, the user's perception will not be of internationalization
and behavior that is culturally friendly. One thing most of us have
almost certainly learned from the experience with email usage is that
users strongly prefer email addresses that closely resemble names or
initials to those involving meaningless strings of letters or
numbers. If the names or initials of the names in the email address
can be expressed in the native languages and writing systems of the
users, the Internet will be perceived as more natural 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
headers, or of permitting upgraded mail user agents (MUAs) to decode
a special coding and display local characters. To be perceived as
usable by end users, the addresses must be internationalized, and
handled consistently, in all of the contexts in which they occur.
That requirement has far-reaching implications: collections of
patches and workarounds are not adequate. Even if they were
adequate, that approach risks an assortment of implementations with
different sets of patches and workarounds having been applied with
consequent user confusion about what is actually be run 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 (see [I18Nemail-
constraints] for an extended discussion of this optimization). 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 headers, and (finally) a requirement for support of
the 8BITMIME option so that all of this can be transported through
the mail system without having to overcome the limitation that
headers do not have content-transfer-encodings.

1.3. Terminology

This document assumes a reasonable understanding of the protocols and
terminology of the core email standards as documented in [RFC2821]
and [RFC2822].

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 "protocols on the wire" principle. That email
architecture, as it has evolved, and 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).

In this document, an address is "all-ASCII" if every character in the
address is in the ASCII character repertoire [ASCII]; an address is
"non-ASCII" if any character is not in the ASCII character
repertoire. 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 term "internationalized email address", or "IMA", refers to an
address permitted by this specification. [[anchor3: Note in Draft/
Placeholder: it appears that the term "IMA" is not used in a precise
and consistent way across the document set. It is sometimes used to
refer simply to a "non-ASCII" address; sometimes to an address that
contains non-ASCII characters, even if that address is encoded into
ASCII characters (i.e., as an ACE); and sometimes as an address that
may contain non-ASCII characters but may also be a traditional
adress. The definition needs to be clarified in an upcoming draft
and all uses of the term brought into line with the definition.]]

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

2. Overview of the Approach

This set of specifications changes both SMTP and the format of email
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 operational and implementation suggestions and guidance for
the protocols.

3. Document Roadmap

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

o SMTP extensions. This document provides an SMTP extension for
   internationalized addresses, as provided for in RFC 2821
   [I18Nemail-SMTPext].
o Email headers in UTF-8. This document essentially updates RFC
   2822 to permit some information in email headers to be expressed
   directly by Unicode characters encoded in UTF-8 when the SMTP
   extension is used [I18Nemail-UTF8].

   o In-transit downgrading from internationalized addressing with the
      SMTP extension and UTF-8 headers to traditional email formats and
      characters [I18Nemail-downgrade]. Downgrading either at the point
      of message origination or after the mail has successfully been
      received by a final delivery SMTP server (sometimes called an
      "MDA") involve different constraints and possibilities; see
      Section 4.3 and Section 5, below.
   o Operational guidelines and suggestions for the deployment of
      internationalized email [I18Nemail-ops].
   o Special considerations for mailing lists and similar distributions
      during the transition to internationalized email [I18Nemail-
      Exploder].
   o Design decisions, history, and alternative models for
      internationalized Internet email [I18Nemail-history].

4. Overview of Protocol Extensions and Changes

4.1. SMTP Extension for Internationalized eMail Address

   An SMTP extension, "IMA" [[anchor7: Extension name should be
   corrected when we make a final decison and synchronized with the
   "I18Nemail-SMTPext" document]] is specified that
   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 headers (see
      the next subsection)
   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.
   o Provides information to support downgrading mechanisms.

   Some general principles apply to this work.
   1. Whatever encoding is used should apply to the whole address and
       be directly compatible with software used at the user interface.
   2. An SMTP relay must
       * Either recognize the format explicitly, agreeing to do so via
          an ESMTP option,
       * Select and use an ASCII-only address, or
       * Bounce the message so that the sender can make another plan.
   3. In the interest of interoperability, charsets other than UTF-8
       are prohibited. There is no practical way to identify them
       properly with an extension similar to this without introducing
       great complexity.

4.2. **Transmission of Email Header in UTF-8 Encoding**

   There are many places in MUAs or in user presentation in which email
   addresses or domain names appear. Examples include the conventional
   From, To, or Cc header fields; Message-IDs; In-Reply-To fields that
   may contain addresses or domain names; in message bodies; or
   elsewhere. We must examine all of them 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 ACE variants, are used, the user
   will inevitably see them, at least occasionally, 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. But 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 nearly as possible the same as the
   encodings used in message headers and message bodies.

   It seems clear that the point at which email local parts are
   internationalized is the point that email headers should simply be
   shifted to a full internationalized form, presumably using UTF-8
   rather than ASCII as the base character set for other than protocol
   elements such as the header field names themselves. The transition
   to that model includes support for address, and address-related,
   fields within the headers of legacy systems. This is done by
   extending the encoding models of [RFC2045] and [RFC2231]. However,
   our target should be fully internationalized headers, as discussed
   [I18Nemail-UTF8].

4.3. **Downgrading Mechanism for Backward Compatibility**

   As with any use of the SMTP extension mechanism, there is always a
   possibility of a client that requires the feature encountering a
   server that does not. In the case of IMA, the risk should be
   minimized by the fact that the selection of submission servers are
   presumably under the control of the sender's client and the selection
   of potential intermediate relays is under the control of the
   administration of the final delivery server.

   For those situations, there are basically two possibilities:
   o Reject or bounce the message, requiring the sender to resubmit it
      with traditional-format addresses and headers.
   o Figure out a way to downgrade the envelope or message body in
      transit. Especially when internationalized addresses are
      involved, downgrading will require either that an all-ASCII
      address be obtained from some source or computed. An optional

extension parameter is provided as a way of transmitting an
alternate address. Computing an all-ASCII form of a non-ASCII
address requires that the sender have some knowledge. This
knowledge is normally restricted to final delivery servers, but
some extensions may be feasible there too. Downgrade issues and a
specification are discussed in [I18Nemail-downgrade].

The first of these two options, that of rejecting or returning the
message to the sender MAY always be chosen.

There is also a third case, one in which the client is I18Nemail-
capable, the server is not, but the message does not require the
extended capabilities. In other words, both the addresses in the
envelope and the entire set of headers of the message are entirely in
ASCII (perhaps including encoded-words in the headers). In that
case, the client SHOULD send the message whether or not the server
announces the IMA capability.

5. Downgrading Before and After SMTP Transactions

In addition to the in-transit downgrades discussed above, downgrading
may also occur before or during initial message submission or after
delivery to the final delivery MTA. Because these cases have a
different set of available information from in-transit cases, the
constraints and opportunities may be somewhat different too. These
two cases are discussed in the subsections below.

5.1. Downgrading Before or During Message Submission

Perhaps obviously, the most convenient time to convert an address or
message from internationalized to conventional ASCII form is at the
originating MUA, either before the message is sent or after the
internationalized form of the message is rejected or bounced by some
MTA in the path to the presumed destination. 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, translating the message 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.

5.2. 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 by client
software via some email retrieval mechanisms such as POP, IMAP or

others.

The SMTP extension described in Section 4.1 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 headers from accessing stored
internationalized emails.

Since the final delivery SMTP server (to be more specific, its
corresponding mail storage agent) cannot safely assume that agents
accessing email storage will be 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 the case, the final delivery SMTP server MUST
include a mechanism preserve the original internationalized forms
without information loss to support access by I18Nemail-aware agents.

The method and format for downgrading at the final delivery SMTP
server is [[anchor9: will be]] discussed in [I18Nemail-imap-pop].

[[anchor10: Note in draft: There are at least four cases. Both MUA
and IMAP/POP are compliant. Both are non compliant. And only of
them is compliant. Do we need to invent different methods for each
case?]]

6. Advice to Designers and Operators of Mail-receiving Systems

[[anchor12: Note in draft: The material that follows contains some
forward-looking, predictive, statements about discussions to occur
and documents to be written. Be sure they are true before Last
Call.]]

In addition to the protocol specification materials in this set of
documents, the working group has had extensive discussions about
operational considerations in the use of internationalized addresses.
Those topics include how such addresses should be chosen, how they
should relate to ASCII alternatives if such alternatives exist, the
management of mailing lists that might support and contain a mixture
of all-ASCII and non-ASCII addresses, and so on. Those issues are
discussed in [I18Nemail-ops] and [I18Nemail-Exploder].

7. Internationalization Considerations

This entire specification addresses issues in internationalization
and especially the boundaries between internationalization and
localization and between network protocols and client/user interface

actions.

8. Additional Issues

This section identifies issues that are not covered as part of this
set of specifications, but that will need to be considered as part of
IMA deployment.

8.1. Impact on IRIs

The mailto: schema in IRI [RFC3987] may need to be modified when IMA
is standardized.

8.2. POP and IMAP

While SMTP takes care of the transportation of messages, IMAP
[RFC3501] and POP3 [RFC1939] are among mechanisms used to handle the
retrieval of mail objects from a mail store by a client. The use of
internationalized mail addresses or UTF-8 headers will require
extensions to POP and IMAP and/or modifications to the design and
implementation of mail stores and the mechanisms that final delivery
SMTP servers use to put mail into them. However, those mechanisms
are separate from those associated with transport across the network
and are not discussed in this series of documents. The general
issues are [[anchor17: will be]] covered in [I18Nemail-imap-pop].
Some preliminary discussion appears in in Section 5.2.

9. 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.

10. 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
one of the proposed fixes for, e.g., URLs, does not work for email
local parts since they are case-sensitive. That fix involves forcing
all elements that are displayed to be in lower-case and normalized.

Since email addresses are often transcribed from business cards and
notes on paper, they are subject to problems arising from confusable
characters. 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 headers must not
leave the Internet less secure than it is that without the required
extensions. The requirements and mechanisms documented in this set
of IMA specifications do not, in general, raise any new security
issues other than those associated with confusable characters -- a
topic that is being explored thoroughly elsewhere [IDN-nextsteps].
Specific issues 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 IMA and
ASCII addresses.

In addition, email addresses are used in many contexts other than
sending mail, such as for identifiers under various circumstances.
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.

11. **Acknowledgements**

This document, and the related ones, were originally derived from
drafts by John Klensin and the JET group [Klensin-emailaddr], [JET-
IMA]. The work drew inspiration from discussions on the "IMAA"
mailing list, sponsored by the Internet Mail Consortium and
especially from an early draft by Paul Hoffman and Adam Costello
[Hoffman-IMAA] that attempted to define an MUA-only solution to the
IMA problem. [[anchor20: Note in draft: may want to move some of this
to "history" or reference it]]

12. **Change History**

[[anchor22: Note to RFC Editor: this section to be removed prior to
publication]]

12.1. **Version 00**

This version supercedes draft-lee-jet-ima-00 and
draft-klensin-emailaddr-i18n-03. It represents a major rewrite and

change of architecture from the former and incorporates many ideas
and some text from the latter.

12.2. Version 01

o Some clarifications of terminology (more to follow) and general
   editorial improvements.
o Upgrades to reflect discussions during IETF 64.
o Improved treatment of downgrading before and after message
   transport.

13. References

13.1. 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.

[I18Nemail-Exploder]
             "Placeholder: whatever we call the mailing list document",
             2005.

             This document is expected to be developed by the WG. The
             date given here is purely arbitrary.

[I18Nemail-SMTPext]
             Yao, J., Ed. and X. Lee, Ed., "SMTP extension for
             internationalized email address", draft-yao-smtpext-01
             (work in progress), January 2006.

[I18Nemail-UTF8]
             Yeh, J., "Transmission of Email Headers in UTF-8
             Encoding", draft-yeh-ima-utf8headers-00 (work in
             progress), October 2005.

[I18Nemail-constraints]
             Klensin, J., "Internationalization in Internet
             Applications: Issues, Tradeoffs, and Email Addresses",
             February 2006.

[I18Nemail-downgrade]
             YONEYA, Y., Ed. and K. Fujiwara, Ed., "Downgrading

                  mechanism for Internationalized eMail Address (IMA)",
                  draft-yoneya-ima-downgrade-00 (work in progress),
                  October 2005.

   [I18Nemail-ops]
                  "Placeholder: whatever we call the operations document",
                  2005.

                  This document is expected to be developed by the WG. The
                  date given here is purely arbitrary.

   [RFC1651] Klensin, J., Freed, N., Rose, M., Stefferud, E., and E.
              Crocker, "SMTP Service Extensions", RFC 1651, July 1994.

   [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'", RFC 2119, March 1997.

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

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

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

13.2. Informative References

   [Hoffman-IMAA]
                  Hoffman, P. and A. Costello, "Internationalizing Mail
                  Addresses in Applications (IMAA)", draft-hoffman-imaa-03
                  (work in progress), October 2003.

   [I18Nemail-history]
                  Klensin, J., "Decisions and Alternatives for
                  Internationalization of Email Addresses", April 2006.

                  This document is expected to be developed by the WG. The
                  date given here is purely arbitrary.

   [I18Nemail-imap-pop]
                  Klensin, J., "Considerations for IMAP and POP in
                  Conjunction with Email Address Internationalization",

draft-klensin-ima-imappop-00a (work in progress),
April 2006.

This document is expected to be developed by the WG. The
date given here is purely arbitrary.

[IDN-nextsteps]
Klensin, J. and P. Faltstrom, "Review and Recommendations
for Internationalized Domain Names (IDN)", February 2006.

[JET-IMA]   Yao, J. and J. Yeh, "Internationalized eMail Address
(IMA)", draft-lee-jet-ima-00 (work in progress),
June 2005.

[Klensin-emailaddr]
Klensin, J., "Internationalization of Email Addresses",
draft-klensin-emailaddr-i18n-03 (work in progress),
July 2005.

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

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

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

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

[RFC2449]   Gellens, R., Newman, C., and L. Lundblade, "POP3 Extension
Mechanism", RFC 2449, November 1998.

[RFC2822]   Resnick, P., "Internet Message Format", RFC 2822,
April 2001.

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

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

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
    MOCOCO, Inc.
    996-1, 11F, Mirae Asset Venture Tower, Daechi-dong
    Gangnam-gu, Seoul 135-280
    Korea

    Email: yw@mrko.pe.kr

Intellectual Property Statement

   The IETF takes no position regarding the validity or scope of any
   Intellectual Property Rights or other rights that might be claimed to
   pertain to the implementation or use of the technology described in
   this document or the extent to which any license under such rights
   might or might not be available; nor does it represent that it has
   made any independent effort to identify any such rights. Information
   on the procedures with respect to rights in RFC documents can be
   found in BCP 78 and BCP 79.

   Copies of IPR disclosures made to the IETF Secretariat and any
   assurances of licenses to be made available, or the result of an
   attempt made to obtain a general license or permission for the use of
   such proprietary rights by implementers or users of this
   specification can be obtained from the IETF on-line IPR repository at
   http://www.ietf.org/ipr.

   The IETF invites any interested party to bring to its attention any
   copyrights, patents or patent applications, or other proprietary
   rights that may cover technology that may be required to implement
   this standard. Please address the information to the IETF at
   ietf-ipr@ietf.org.

Disclaimer of Validity

   This document and the information contained herein are provided on an
   "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS
   OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET
   ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED,
   INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE
   INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED
   WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Statement

   Copyright (C) The Internet Society (2006). This document is subject
   to the rights, licenses and restrictions contained in BCP 78, and
   except as set forth therein, the authors retain all their rights.

Acknowledgment

   Funding for the RFC Editor function is currently provided by the
   Internet Society.