Internationalizing Domain Names in Applications (IDNA) version 2
draft-hoffman-idna2-01.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the
provisions of BCP 78 and BCP 79. This document may contain material
from IETF Documents or IETF Contributions published or made publicly
available before November 10, 2008. The person(s) controlling the
copyright in some of this material may not have granted the IETF
Trust the right to allow modifications of such material outside the
IETF Standards Process. Without obtaining an adequate license from
the person(s) controlling the copyright in such materials, this
document may not be modified outside the IETF Standards Process, and
derivative works of it may not be created outside the IETF Standards
Process, except to format it for publication as an RFC or to
translate it into languages other than English.

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

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

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

The list of Internet-Draft Shadow Directories can be accessed at

This Internet-Draft will expire on August 30, 2009.

Copyright Notice

Copyright (c) 2009 IETF Trust and the persons identified as the
document authors. All rights reserved.
Abstract

IDNA has been a world-wide success since it was introduced over five years ago. However, it has some notable deficiencies, including being tied to an old version of the Unicode standard and needless restrictions that prevented some languages from being used. This document describes IDNA version 2, which rectifies those problems while making the fewest changes necessary to the original protocol.

Table of Contents

1. Introduction ................................................. 3
   1.1. Acknowledgements ...................................... 4
   1.2. Conventions Used In This Document ...................... 4

2. Changes to RFC 3490 (IDNA v.1) ............................ 4
3. Changes to RFC 3454 (Stringprep) ............................ 4
4. Changes to RFC 3491 (Nameprep) ............................ 6
5. Changes to RFC 3492 (Punycode) ............................ 6
6. Suggestions for Registries .................................. 7
7. IANA Considerations .......................................... 7
8. Security Considerations ...................................... 7
9. References ................................................... 7
   9.1. Normative References .................................... 7
   9.2. Informative References .................................. 8

Appendix A. Work Still to be Done .............................. 8
Appendix B. Changes between versions ........................... 8
   B.1. Changes between the -00 and -01 drafts ................. 8

Author’s Address ................................................. 9
1. Introduction

This document describes Internationalizing Domain Names in Applications (IDNA) version 2 (hereafter called "IDNAv2"), a direct update to IDNA (hereafter called "IDNAv1"). IDNAv1 consists of four RFCs:

- [RFC3490], "Internationalizing Domain Names in Applications (IDNA)", is the main definition of IDNAv1. This defines the processing rules for IDNA and gives the background for how IDNA works.
- [RFC3454], "Preparation of Internationalized Strings ("stringprep")", defines the general framework for processing non-ASCII strings that are used in IDNA.
- [RFC3491], "Nameprep: A Stringprep Profile for Internationalized Domain Names (IDN)", is a short profile of the rules from the stringprep framework.
- [RFC3492], "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)", defines the encoding used in IDNAv1 labels.

Any legal IDNAv1 label that had only visible characters has exactly the same representation in IDNAv2. New labels are allowed in IDNAv2 that were not allowed in IDNAv1.

IDNA needs to be updated for many reasons, some of which are covered in [RFC4690]. If for no other reason, many characters that could appear in domain names have been added since Unicode version 3.2 [UNICODE32], which is the version of the Unicode Standard on which IDNAv1 is based.

One explicit goal of this update is to allow labels with characters that have been added since Unicode version 3.2 to be used in IDNA. To that end, IDNAv2 is based on Unicode 5.1 [UNICODE51]. The tables in stringprep and Nameprep are updated to reflect this change.

Another explicit goal of this update is to not change the encoding of any label of visible characters that is legal in IDNAv1. If an internationalized label of visible characters in IDNAv1 produces an ACE label, IDNAv2 must produce the same ACE label. If an internationalized label of visible characters in IDNAv1 produces an ASCII label, IDNAv2 must produce the same ASCII label. IDNAv2 changes the mapping of two non-visible characters that are common in Arabic and Persian languages from being "mapped to nothing" to being encoded characters.

A third explicit goal is to update the bidirectional ("bidi") algorithm used by IDNAv1 to cover more languages such as Dhivehi and Yiddish. This is done to cover an oversight in IDNAv1 that was...
discovered after the work was finished.

1.1. Acknowledgements

The first serious work on updating IDNAv1 was undertaken by John Klensin, Patrik Faltstrom, Harald Alvestrand, and Cary Karp. It led to the formation of the IDNabis Working Group in the IETF, and they produced many revisions of their documents in that WG. Some of the ideas in this IDNAv2 document (most notably, the update to the bidi algorithm) is derived from their efforts.

Many, many people worked on IDNAv1. In addition to the authors of the standards (Marc Blanchet, Adam Costello, Patrik Faltstrom, and me), there were literally dozens of active participants in the original IDN Working Group in the IETF that began in 2000. Their tireless effort led to IDNAv1.

1.2. Conventions Used In This Document

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 [RFC2119].

In sections of this document where changes are made to RFCs, those changes are shown with a vertical line character ("|") in the first column.

2. Changes to RFC 3490 (IDNA v.1)

All references to the Unicode Standard are updated to refer to [UNICODE51].

All references to Nameprep are updated to refer to the Nameprep in this document. Similarly, all references to stringprep are updated to refer to the stringprep in this document.

In section 3.1, the first bullet point ("1) Whenever dots are used...") is changed to add the following at the end of the sentence: "U+2CFE (Coptic full stop)".

3. Changes to RFC 3454 (Stringprep)

[[[ -----------------------------------------------

NOTE FOR EARLY VERSIONS OF THIS DRAFT

]]]
This section is intentionally incomplete. The tables in Stringprep need to be added to based on the characters added to the repertoire after Unicode 3.2 up to and including Unicode 5.1.

Probably the best way for this to be done is a few dedicated individuals go through the new characters one-by-one, and also to go through them programmatically, and see which tables need to be added to. I have done a first pass of doing this one-by-one, but I felt that publishing my results in the first draft would cause others to get lazy about this important task. Future versions of this document will reflect the results of that work.

The character review will be similar to what we did in IDNAv1, except that we don’t have to create any new buckets. Basically, we have to see whether a particular new character should be mapped to nothing, or whether it should be prohibited for one of the reasons already listed in RFC 3454. In my not-careful first pass, I found very few characters that will need to be added to sections 3 or 5. The case-mapping will happen algorithmically, with a check that the new map does not change any value in the old map.

============================================================ 

Most of the changes to RFC 3454 are to add characters to the tables in the document. These characters come from Unicode version 5.1. Thus, the tables become valid for Unicode version 5.1. However, the same tables are still valid for Unicode version 3.2 because a profile that is still using version 3.2 will not ever use the added rows in the updated tables.

In all places other than Appendix A, references to "[Unicode3.2]" are updated to refer to [UNICODE51]. Similarly, all text references to "Unicode version 3.2" are updated to "Unicode version 5.1".

Characters will be added to the tables in section 3.1 to reflect the differences between Unicode 3.2 and Unicode 5.1. For example, U+E0100 to U+E01EF will be added to the second list in the section.

In section 3.2, in order to allow for much more correct encoding of Arabic and Persian scripts, remove the following from the table of characters mapped to nothing:

200C; ZERO WIDTH NON-JOINER
200D; ZERO WIDTH JOINER

In section 3.2, change "CaseFolding-3.txt" to "CaseFolding.txt".

Characters will be added to the tables in subsections of section 5.
An example is that U+2064 will be added to the list in section 5.2.

In section 6, at the end of the fourth paragraph (which currently ends with "have bidirectional category "EN"."), the following sentence is added: "The Unicode Standard also defines a bidirectional category "NSM" for "non-spacing marks"."

In section 6, the third requirement is changed to read:

| 3) If a string contains any RandALCat character, the first character MUST be a RandALCat character, and the last characters of the string must be either a RandALCat character or a RandALCat character followed by one or more NSM characters. |

In the references, update the reference for UAX15, and add a reference for [UNICODE51].

Appendix A is changed to read:

| The following is the only repertoire covered in this document: |
| - Unicode 3.2, as defined in [UNICODE32] |
| - Unicode 5.1, as defined in [UNICODE51] |

A new appendix, "A.2 Unassigned code points in Unicode 5.1", will be added.

The tables in appendixes B, C, and D will be added to.

4. Changes to RFC 3491 (Nameprep)

All references to IDNA and stringprep are updated to refer to the stringprep in this document.

In section 1 and 2, "Unicode 3.2" is changed to "Unicode 5.1".

In section 10, change the last table entry to "This is the second version of Nameprep."

5. Changes to RFC 3492 (Punycode)

IDNAv2 does not change RFC 3492.
6. Suggestions for Registries

This is a placeholder for a short section that covers new advice for registries that was not included in IDNAv1. It will include ideas about multi-script labels, not allowing registration that includes ZWNJ and ZWJ in labels that do not need them, and possibly other advice.

7. IANA Considerations

IANA is requested to add the following to the stringprep profile registry (www.iana.org/assignments/stringprep-profiles).

Name of this profile: Nameprep

RFC in which the profile is defined: This document.

Indicator whether or not this is the newest version of the profile: This is the second version of Nameprep.

8. Security Considerations

The security considerations from RFCs 3454, 3490, 3491, and 3492 all apply to this document. The changes between IDNAv1 and IDNAv2 are not believed to add any new security considerations.

9. References

9.1. Normative References


9.2. Informative References


Appendix A. Work Still to be Done

Figure out exactly how we want the reference to Unicode 3.2 and Unicode 5.1 to look in the references section, then figure out how to wrestle xml2rfc to produce that.

Fill in all the tables for the updates to stringprep.

Decide if this entire document should be about Unicode 5.2, which is expected out by mid-2009.

Appendix B. Changes between versions

(This section is to be removed by the RFC Editor.)

B.1. Changes between the -00 and -01 drafts

In section 1, changed the target for backwards-compatibility to be for strings that have only visible characters.

In section 3, removed the first paragraph.

In section 3 (about Stringprep section 3.1), added the text about removing U+200C and U+200D from the mapped-to-nothing list.

In section 3 (about Stringprep section 6), replaced:
3) If a string contains any RandALCat character, a RandALCat character MUST be the first character of the string, and either a RandALCat character or NSM character MUST be the last character of the string.

with

3) If a string contains any RandALCat character, the first character MUST be a RandALCat character, and the last characters of the string must be either a RandALCat character or a RandALCat character followed by one or more NSM characters.

Added new placeholder section 6 on advice to registries.

In Appendix A, added the thought about targeting Unicode 5.2 instead of Unicode 5.1.

Author’s Address

Paul Hoffman

Email: phoffman@imc.org