Network Working Group
Internet-Draft
Updates: 5890, 5891 (if approved)
Intended status: Standards Track
Expires: September 12, 2017

J. Klensin
A. Freytag
ASMUS, Inc.

March 11, 2017

# Internationalized Domain Names in Applications (IDNA): Registry Restrictions and Recommendations
# draft-klensin-idna-rfc5891bis-00.txt

Abstract

The IDNA specifications for internationalized domain names combine rules that determine the labels that are allowed in the DNS without violating the protocol itself and an assignment of responsibility, consistent with earlier specifications, for determining the labels that are allowed in particular zones. Conformance to IDNA by registries and other implementations requires both parts. Experience strongly suggests that the language describing those responsibility was insufficiently clear to promote safe and interoperable use of the specifications and that more details and some specific examples would have been helpful. This specification updates the earlier ones to provide that guidance and to correct some technical errors in the descriptions. It does not alter the protocols and rules themselves in any way.

Status of This Memo

Table of Contents



## 1. Introduction

Parts of the specifications for Internationalized Domain Names in Applications (IDNA) [RFC5890] [RFC5891] [RFC5894] (collectively known, along with RFC 5892 [RFC5892], RFC 5893 [RFC5893] and updates to them, as "IDNA2008" (or just "IDNA") impose a requirement that domain name system (DNS) registries restrict the characters they allow in domain name labels (see Section 2 below), and the contents and structure of those labels. That requirement and restriction are consistent with the "trustee for the community" requirements of the original specification for DNS naming and authority [RFC1591]. The restrictions are intended to limit the permitted characters and strings to those for which the registries or their advisers have a thorough understanding and for which they are willing to take responsibility.

That provision is centrally important because it recognized that historical relationships and variations among scripts and writing systems, the continuing evolution of those systems, differences in the uses of characters among languages (and locations) that use the same script, and so on make it impossible for a single list of characters and simple rules to be able to generate an "if we use these, we will be safe from confusion and various attacks" guideline.

Instead, the algorithm and rules of RFC 5981 and 5982 eliminate many of the most dangerous and otherwise problematic cases, but cannot eliminate the need for registries and registrars to understand what they are doing and taking responsibility for the decisions they make.

The way in which the IDNA2008 specifications expressed these requirements may have obscured the intention that they actually are requirements. Section 2.3.2.3 of the Definitions document [RFC5890] mentions the need for the restrictions, indicates that they are mandatory, and points the reader to section 4.3 of the Protocol document [RFC5891], which in turn points to Section 3.2 of the Rationale document [RFC5894], with each document providing further detail, discussion, and clarification.

This specification is intended to unify and clarify these requirements for registry decisions and responsibility and to emphasize the importance of registry restrictions at all levels of the DNS. It also makes a specific recommendation for character repertoire subsetting intermediate between the code points allowed by RFC 5891 and 5892 and those allowed by individual registries. It does not alter the basic IDNA2008 protocols and rules themselves in any way.

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 RFC 2119 [RFC2119].

[[NOTE IN DRAFT: While the co-authors have gone through several iterations of this I-D and discussed some sectons of it with others, it is still a very preliminary version. In particular, the specific material in Section 4 has not yet been inserted.]]

## 2. Registry Restrictions in IDNA2008

As mentioned above, IDNA2008 specifies that the registries for each zone in the DNS that supports IDN labels are required to develop and apply their own rules to restrict the allowable labels, including limiting characters they allow to be used in labels in that zone. The chosen list MUST BE smaller than the collection of code points specified as "PVALID", "CONTEXTJ", and "CONTEXTO" by the rules

established by the protocols themselves. The latter two categories, and labels containing any characters that are normally part of a script written right to left [RFC5893], require that additional rules, specified in the protocols and known as "contextual rules" and "bidi rules", be applied. The entire collection of rules and restrictions required by the IDNA2008 protocols themselves are known as "protocol restrictions".

As mentioned above, registries may apply (and generally are required to apply) additional rules to further restrict the list of permitted code points, contextual rules (perhaps applied to normally PVALID code points) that apply additional restrictions, and/or restrictions on labels. The most obvious of those restrictions include provisions for restricting suggested new registrations based on conflicts with labels already registered in the zone and specifications of what constitutes such conflicts based on the properties of the labels in question. They further include prohibitions on code points and labels that are not consistent with the intended function of the zone or the subtree in which it is embedded (see Section 3) or limitations on where in a label allowable code points may be placed.

These per-registry (or per-zone) rules are commonly known as "registry restrictions" to distinguish them from the protocol restrictions described above. By necessity, the latter are somewhat generic, having to cater both to the union of the needs for all zones, as well as to the most permissive zones. In consequence, additional Registry restrictions are essential to provide for the necessary security in the face of the tremendous variations and differences in writing systems, their ongoing evolution and development, as well as the human ability to recognize and distinguish characters in different scripts around the world and under different circumstances.

## 3. Progressive Subsets of Allowed Characters

The algorithm and rules of RFC 5891 and 5892 set an absolute upper bound on the code points that can be used in domain name labels; registries MUST NOT include code points unless they are allowed by those rules. Each registry that intends to allow IDN registrations MUST then determine which code points will be allowed by that registry and SHOULD consider additional rules, including contextual and whole label restrictions that provide further protection for registrants and users. For example, the widely-used principle that bars labels containing characters from more than one script is not an IDNA2008 requirement. It has been adopted by many registries but, as Section 4.4 of RFC 5890 indicates, there may be circumstances in which is it not required or appropriate.

In formulating their own rules, registries SHOULD normally consult carefully-developed consensus recommendations about global maximum repertoires to be used such as the ICANN Maximal Starting Repertoire 2 (MSR-2) for the Development of Label Generation Rules for the Root Zone [ICANN-MSR2] (or its successor documents). Additional recommendations of similar quality about particular scripts or languages exist, including, but not limited to, the RFCs for Cyrillic [RFC5992] or Arabic Language [RFC5564] or script-based repertoires from the approved ICANN Root Zone Label Generation Rules (LGR-1) [ICANN-LGR1] (or its successor documents).

It is the responsibility of the registry to determine which, if any, of those recommendations are applicable and to further subset or extend them as needed. For example, several of the recommendations are designed for the root zone and therefore exclude digits and U+002D HYPHEN-MINUS, a restriction not generally appropriate for other zones. On the other hand, some zones may be designed to not cater for all users of a given script, but perhaps only for the needs of selected languages, in which case a more selective repertoire may be appropriate.

In making these determinations, a registry SHOULD follow the IAB guidance in RFC 6912 [RFC6912]. Those guidelines include a number of principles for use in making decisions about allowable code points. In addition, that document notes that the closer a particular zone is to the root, the more restrictive the space of permitted labels should be. RFC 5894 provides some suggestions for any registry that may decide to reduce opportunities for confusion or attacks by constructing policies that disallow characters used in historic writing systems (whether these be archaic scripts or extensions of modern scripts for historic or obsolete orthographies) or characters whose use is restricted to specialized, or highly technical contexts. These suggestions were among the principles guiding the design of ICANN's Maximal Starting Repertoires [LGR-Procedure].

Particularly for a zone for which all labels to be delegated are not for the use of the same organization or enterprise, a registry decision to allow only those code points in the full repertoire of the MSR (plus digits and hyphen) would already avoid a number of issues inherent in a more permissive policy like "use anything permitted by IDNA2008", while still supporting the native languages and scripts for the vast majority of users today. However, it is unlikely, by itself, to fully satisfy the mandate set out above for three reasons.

1. The MSR, like the set of code points permissible under IDNA2008 itself, was conceived merely as an upper bound on permissible letter code points (it excludes digits and the hyphen). It was

always intended to be used as a starting point for setting registry policy, with the expectation that some of the code points in the MSR would not be included in the final registry policy, whether for lack of actual usage, or for being inherently problematic.

2. It was recognized that many scripts require contextual rules for many more code points than are covered by CONTEXTO or CONTEXTJ rules defined in IDNA2008. This is particularly true for combining marks, typically used to encode diacritics, tone marks, vowel signs and the like. While, theoretically, any combining mark may occur in any context in Unicode, in practice rendering and other software that users rely on in viewing or entering labels will not support arbitrary combining sequences, or indeed arbitrary combinations of code points, in the case of complex scripts.

   Contextual rules are required to limit allowable code point sequences to those that can be expected to be rendered reliably. Identifying those requires knowledge about the way code points are used in a script, whence the mandate for registries to only support code points they understand. In this, some of the other recommendations, such as the Informational RFCs for specific scripts (e.g., Cyrillic [RFC5992]) or languages (e.g., Arabic [RFC5564] or Chinese [RFC4713]), or the Root Zone LGRs developed by ICANN, may provide useful guidance.

3. Third, because of the widely accepted practice of limiting any given label to a single script, a universal repertoire, such as the MSR, would have to be divided on a per script basis into subrepertoires to make it useful, with some of those repertoires overlapping, for example, in the case of East Asian shared usage of the Han ideographs.

Registries choosing to make exceptions and allow code points that recommendations such as the MSR do not allow should make such decisions only with great care and only if they have considerable understanding of, and great confidence in, their appropriateness. The obvious exception from the MSR would be to allow digits and the hyphen. Neither were allowed by the MSR, but only because they are not allowed in the Root Zone.

Nothing in this document permits a registry to allow code points or labels that are disallowed or otherwise prohibited by IDNA2008.

## 4. Other corrections and updates

After the initial IDNA2008 documents were published (and RFC 5892 was updated for Unicode 6.0 by RFC 6452 [RFC6452]) several errors or instances of confusing text were noted. For the convenience of the community, the relevant corrections for RFC 5890 and 5891 are noted below and update the corresponding documents. There are no errata for RFC 5893 or 5894 as of the date this document was published. Because further updates to RFC 5892 would require addressing other pending issues, the outstanding erratum for that document is not considered here. For consistency with the original documents, references to Unicode 5.0 are preserved.

### 4.1. Updates to RFC 5890

Errata ID 4695: The maximum length mess
... to be supplied ...

Errata ID 4824: More comments about length
Note: "Hold for doc update".
... to be supplied ...

Errata ID 4823: More comments about length
Note: "Hold for doc update".
... to be supplied ...

### 4.2. Updates to RFC 5891

Errata ID 3969: Improve reference for combining marks
Note: "Hold for doc update".

## 5. Security Considerations

As discussed in IAB recommendations about internationalized domain names [RFC4690], [RFC6912], and elsewhere, poor choices of strings for DNS labels can lead to opportunities for attacks, user confusion, and other issues less directly related to security. This document clarifies the importance of registries carefully establishing design policies for the labels they will allow and that having such policies and taking responsibility for them is a requirement, not an option. If that clarification is useful in practice, the result should be an improvement in security.

## 6. Acknowledgements

... placeholder ...

## 7. IANA Considerations

[[CREF1: RFC Editor: Please remove this section before publication.]]

This memo includes no requests to or actions for IANA. In particular, it does not contain any provisions that would alter any IDNA-related registries or tables.

## 8. References

### 8.1. Normative References

[ICANN-LGR1] ICANN, "Root Zone Label Generation Rules (LGR-1)", June 2015, <https://www.icann.org/resources/pages/root-zone-lgr-2015-06-21-en>.

[ICANN-MSR2] ICANN, "Maximal Starting Repertoire Version 2 (MSR-2) for the Development of Label Generation Rules for the Root Zone", April 2015, <https://www.icann.org/news/announcement-2-2015-04-27-en>.

[RFC1591] Postel, J., "Domain Name System Structure and Delegation", RFC 1591, DOI 10.17487/RFC1591, March 1994, <http://www.rfc-editor.org/info/rfc1591>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, DOI 10.17487/RFC5890, August 2010, <http://www.rfc-editor.org/info/rfc5890>.

[RFC5891] Klensin, J., "Internationalized Domain Names in Applications (IDNA): Protocol", RFC 5891, DOI 10.17487/RFC5891, August 2010, <http://www.rfc-editor.org/info/rfc5891>.

[RFC5892Erratum] "RFC5892, "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", August 2010, Errata ID: 3312", Errata ID 3312, August 2012, <http://www.rfc-editor.org/errata_search.php/doc/html/rfc5892>.

[RFC5894] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Background, Explanation, and Rationale", RFC 5894, DOI 10.17487/RFC5894, August 2010, <http://www.rfc-editor.org/info/rfc5894>.

## 8.2. Informative References

[LGR-Procedure]
Internet Corporation for Assigned Names and Numbers (ICANN), "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels", March 2013, <https://www.icann.org/en/system/files/files/draft-lgr-procedure-20mar13-en.pdf>.

[RFC4690] Klensin, J., Faltstrom, P., Karp, C., and IAB, "Review and Recommendations for Internationalized Domain Names (IDNs)", RFC 4690, DOI 10.17487/RFC4690, September 2006, <http://www.rfc-editor.org/info/rfc4690>.

[RFC4713] Lee, X., Mao, W., Chen, E., Hsu, N., and J. Klensin, "Registration and Administration Recommendations for Chinese Domain Names", RFC 4713, DOI 10.17487/RFC4713, October 2006, <http://www.rfc-editor.org/info/rfc4713>.

[RFC5564] El-Sherbiny, A., Farah, M., Oueichek, I., and A. Al-Zoman, "Linguistic Guidelines for the Use of the Arabic Language in Internet Domains", RFC 5564, DOI 10.17487/RFC5564, February 2010, <http://www.rfc-editor.org/info/rfc5564>.

[RFC5892] Faltstrom, P., Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, DOI 10.17487/RFC5892, August 2010, <http://www.rfc-editor.org/info/rfc5892>.

[RFC5893] Alvestrand, H., Ed. and C. Karp, "Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)", RFC 5893, DOI 10.17487/RFC5893, August 2010, <http://www.rfc-editor.org/info/rfc5893>.

[RFC5992] Sharikov, S., Miloshevic, D., and J. Klensin, "Internationalized Domain Names Registration and Administration Guidelines for European Languages Using Cyrillic", RFC 5992, DOI 10.17487/RFC5992, October 2010, <http://www.rfc-editor.org/info/rfc5992>.

[RFC6452] Faltstrom, P., Ed. and P. Hoffman, Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA) - Unicode 6.0", RFC 6452, DOI 10.17487/RFC6452, November 2011, <http://www.rfc-editor.org/info/rfc6452>.

[RFC6912] Sullivan, A., Thaler, D., Klensin, J., and O. Kolkman, "Principles for Unicode Code Point Inclusion in Labels in the DNS", RFC 6912, DOI 10.17487/RFC6912, April 2013, <http://www.rfc-editor.org/info/rfc6912>.

## Authors' Addresses

John C Klensin
1770 Massachusetts Ave, Ste 322
Cambridge, MA 02140
USA

Phone: +1 617 245 1457
Email: john-ietf@jck.com

Asmus Freytag
ASMUS, Inc.

Email: asmus@unicode.org