Network Working Group
Internet-Draft
Expires: December 30, 2002

M. Duerst
W3C/Keio University
July 1, 2002

Internationalized Domain Names in URIs
draft-ietf-idn-uri-02

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

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 http://www.ietf.org/shadow.html.

This Internet-Draft will expire on December 30, 2002.

Copyright Notice

Abstract

This document proposes to upgrade the definition of URIs (RFC 2396) [RFC2396] to work consistently with internationalized domain names.

Table of Contents

1. Introduction

Internet domain names serve to identify hosts and services on the Internet in a convenient way. The IETF IDN working group [IDNWG] has been working on extending the character repertoire usable in domain names beyond a subset of US-ASCII.

One of the most important places where domain names appear are Uniform Resource Identifiers (URIs, [RFC2396], as modified by [RFC2732]). However, in the current definition of the generic URI syntax, the restrictions on domain names are 'hard-coded'. In Section 2, this document relaxes these restrictions by updating the syntax, and defines how internationalized domain names are encoded in URIs.

The syntax in this document has been choosen to further increase the uniformity of URI syntax, which is a very important principle of URIs.

In practice, escaped domanin names should be used as rarely as possible. Wherever possible, the actual characters in Internationalized Domain Names should be preserved as long as possible by using IRIs [IRI] rather than URIs, and only converting to URIs and then to ACE-encoded [IDNA] domain names (or ideally directly to ACE-encoding without even using URIs) when resolving the IRI. Also, this document does in no way exclude the use of ACE encoding directly in an URI domain name part. ACE encoding may be used directly in an URI domain name part if this is considered necessary for interoperability.

Please note that even with the definition of URIs in [RFC2396], some URIs can already contain host names with escaped characters. For example, mailto:example@w%33.org is legal per [RFC2396] because the mailto: URI scheme does not follow the generic syntax of [RFC2396].

2. URI syntax changes

The syntax of URIs [RFC2396] currently contains the following rules relevant to domain names:

```
hostname      = *( domainlabel "." ) toplabel [ "." ]
domainlabel   = alphanum | alphanum *( alphanum | "-" ) alphanum
toplabel      = alpha | alpha *( alphanum | "-" ) alphanum
```

The later two rules are changed as follows:

```
domainlabel   = anchar | anchar *( anchar | "-" ) anchar
toplabel      = achar | achar *( anchar | "-" ) anchar
```

and the following rules are added:

```
anchar        = alphanum | escaped
achar         = alpha | escaped
```

Characters outside the repertoire (alphanum) are encoded by first encoding the characters in UTF-8 [RFC 2279], resulting in a sequence of octets, and then escaping these octets according to the rules defined in [RFC2396].

Using UTF-8 assures that this encoding interoperates with IRIs [IRI]. It is also aligned with the recommendations in [RFC2277] and [RFC2718], and is consistent with the URN syntax [RFC2141] as well as recent URL scheme definitions that define encodings of non-ASCII characters based on UTF-8 (e.g., IMAP URLs [RFC2192] and POP URLs [RFC2384]).

The above syntax rules permit for domain names that are neither permitted as US-ASCII only domain names nor as internationalized domain names. However, such syntax should never be used, and will always be rejected by resolvers. For US-ASCII only domain names, the syntax rules in [RFC2396] are relevant. For example, http://www.w%33.org is legal, because the corresponding 'w3' is a legal 'domainlabel' according to [RFC2396]. However, http://%2a.example.org is illegal because the corresponding '*' is not a legal 'domainlabel' according to [RFC2396]. For domain names containing non-ASCII characters, the legal domain names are those for which the ToASCII operation ([IDNA], [Nameprep]; using the unescaped UTF-8 values as input) is successful.

For consistency in comparison operations and for interoperability with older software, the following should be noted: 1) US-ASCII characters in domain names should not be escaped. 2) Because of the principle of syntax uniformity for URIs, it is always more prudent to take into account the possibility that US-ASCII characters are escaped.

The work of the IDN WG includes some procedures for name preparation [Nameprep]. Before encoding an internationalized domain name in an URI, this preparation step SHOULD be applied. However, the URI resolver MUST also apply any steps required as part of domain name resolution by [IDNA].

3. Security considerations

The security considerations of [RFC2396] and those applying to internationalized domain names apply. There may be an increased potential to smuggle escaped US-ASCII-based domain names across firewalls, although because of the uniform syntax principle for URIs, such a potential is already existing.

4. Change Log

4.1 Changes from draft-ietf-idn-uri--01 to draft-ietf-idn-uri-02

Moved change log to back

Changed to only change URIs; IRI syntax updated directly in IRI draft.

Removed syntax restriction on %hh in the US-ASCII part, but made clear that restrictions to domain names apply.

Made clear that escaped domain names in URIs should only be an intermediate representation.

Gave example of mailto: as already allowing escaped host names.

4.2 Changes from draft-ietf-idn-uri--00 to draft-ietf-idn-uri-01

Changed requirement for URI/IRI resolvers from MUST to SHOULD

Changed IRI syntax slightly (ichar -> idchar, based on changes in [IRI])

Various wording changes

References

[IDNA] Faltstrom, P., Hoffman, P. and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", draft-ietf-idn-idna-09.txt (work in progress), May 2002, <http://www.ietf.org/internet-drafts/draft-ietf-idn-idna-09.txt>.

[IDNWG] "IETF Internationalized Domain Name (idn) Working Group".

[IRI] Duerst, M. and M. Suignard, "Internationalized Resource Identifiers (IRI)", draft-duerst-iri-01 (work in progress), July 2002.

[ISO10646] International Organization for Standardization, "Information Technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane", ISO Standard 10646-1, October 2000.

[Nameprep] Hoffman, P. and M. Blanchet, "Nameprep: A Stringprep Profile for Internationalized Domain Names", draft-ietf-idn-nameprep-10.txt (work in progress), May 2002, <http://www.ietf.org/internet-drafts/draft-ietf-idn-nameprep-10.txt>.

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

[RFC2141] Moats, R., "URN Syntax", RFC 2141, May 1997.

[RFC2192] Newman, C., "IMAP URL Scheme", RFC 2192, September 1997.

[RFC2277] Alvestrand, H., "IETF Policy on Character Sets and Languages", BCP 18, RFC 2277, January 1998.

[RFC2279] Yergeau, F., "UTF-8, a transformation format of ISO 10646", RFC 2279, January 1998.

[RFC2384] Gellens, R., "POP URL Scheme", RFC 2384, August 1998.

[RFC2396] Berners-Lee, T., Fielding, R. and L. Masinter, "Uniform Resource Identifiers (URI): Generic Syntax", RFC 2396, August 1998.

[RFC2640] Curtin, B., "Internationalization of the File Transfer Protocol", RFC 2640, July 1999.

[RFC2718] Masinter, L., Alvestrand, H., Zigmond, D. and R. Petke, "Guidelines for new URL Schemes", RFC 2718, November 1999.

[RFC2732] Hinden, R., Carpenter, B. and L. Masinter, "Format for Literal IPv6 Addresses in URL's", RFC 2732, December 1999.

Author's Address

Martin Duerst
W3C/Keio University
5322 Endo
Fujisawa 252-8520
Japan

Phone: +81 466 49 1170
Fax: +81 466 49 1171
EMail: duerst@w3.org
URI: http://www.w3.org/People/D%C3%BCrst/

Full Copyright Statement

Acknowledgement

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