INTERNET-DRAFT
draft-ietf-idn-uri-00
Expires July 2001

Martin Duerst
W3C/Keio University

January 6, 2001

Internationalized Domain Names in URIs and IRIs

Abstract

This document is a first draft for the provisions necessary to upgrade the definitions of URIs [RFC 2396] and IRIs (Internationalized Resource Identifiers, [IRI]) to work with internationalized domain names.

1. Introduction

Internet domain names serve to identify hosts and services on the Internet in a convenient way. The IETF IDN working group is currently 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, [RFC 2396], as modified by [RFC2732]). However, in the current definition of the generic URI syntax, the restrictions on domain names are 'hard-coded'. This document proposes to relax these restrictions by updating the syntax, and defines how internationalized domain names are encoded in URIs.

URIs themselves are restricted to a subset of US-ASCII. However, there is a proposal for relieving these restrictions by creating a new protocol element called an IRI (Internationalized Resource Identifier [IRI]). While IRIs in general allow the use of non-ASCII characters, the syntax of IRIs has the same restriction for domain names as the syntaxt of URIs. This document proposes to relax these restrictions, too, in a way that is compatible with the new syntax for URIs. This means that encoding an internationalized domain name in an URI and encoding the same name in an IRI will produce an URI and an IRI that can be converted into each other using the procedures defined in [IRI] for these conversions.

2. URI syntax changes

The syntax of URIs [RFC2326] 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   = escalphanum | escalphanum *( escalphanum | "-" )
                       escalphanum
       toplabel      = escalpha | escalpha *( escalphanum | "-" )
                       escalphanum
```

and the following rules are added:

```
       escalphanum   = escaped8 | alphanum
       escalpha      = elcaped8 | alpha
       escaped8      = "%" hexdig8 HEXDIG
       hexdig8       = <<HEXDIG greater than 7>>
```

The %HH escaping is used to encode characters outside the repertoire of US-ASCII. This is done by first encoding the characters in UTF-8 [RFC 2279], resulting in a sequence of octets, and then escaping these octets.

Using UTF-8 assures that this encoding interoperates with IRIs (see Section 3). It is also alligned with the recommendations in [RFC 2277] and [RFC 2718], and is consistent with the URN syntax [RFC2141] as well as recent URL scheme definitions that define encodings of non-ASCII characters based on (e.g., IMAP URLs [RFC 2192] and POP URLs [RFC 2384]).

Please note that the use of UTF-8 for encoding internationalized domain names in URIs is independent of the choice of encoding chosen for these names in the DNS protocol. In case something else than UTF-8 is chosen for the later, a future version of this document may give instructions for the conversion if deemed necessary.

The above syntax rules do not extend the possible domain names based on US-ASCII characters. This may have to be changed in case the IDN WG should decide to allow such extensions.

The above rules also do not allow escaping of US-ASCII characters, although this is allowed in the other parts of an URI (except for the special provisions in case of reserved characters). Allowing such escaping would make the syntax rules quite a bit more complicated, would mean that the restrictions on US-ASCII characters can be circumvented by using escaping, or would lead to much simpler syntax rules that don't express these restrictions anymore. Even in case escaping of US-ASCII characters is allowed in order to simplify processing, it should be noted that it is always better not to escape US-ASCII characters in domain names because of the possibility that a resolver cannot unescape them. At least purely US-ASCII domain names would then always be resolved by such a processor.

While only the restrictions on US-ASCII characters are expressed in the rules above, all the other restrictions on internationalized domain names that will be defined by the IDN WG MUST be respected.

The work of the IDN WG currently includes some procedures for name preparation. Before encoding an internationalized domain name in an URI, this preparation step SHOULD be applied. However, the resolver MUST also apply name preparation.

2. IRI syntax changes

The syntax of IRIs [IRI] 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   = intalphanum | intalphanum *( intalphanum | "-" )
                       intalphanum
       toplabel      = intalpha | intalpha *( intalphanum | "-" )
                       intalphanum
```

and the following rules are added:

```
       intalphanum   = ichar | alphanum | escaped8
       intalpha      = ichar | alpha | escaped8
       escaped8      = "%" hexdig8 HEXDIG
       hexdig8       = <<HEXDIG greater than 7>>
```

where ichar, as in [IRI], is:

```
       ichar         =  << any character of UCS [ISO10646] beyond
```

U+0080, subject to limitations in Section 3.1. of [IRI] >>

With respect to the allowed domain names based on US-ASCII characters, the same considerations as in Section 2 apply.

As in Section 2, all the other restrictions on internationalized domain names that will be defined by the IDN WG MUST be respected. Also, before encoding an internationalized domain name in an IRI, name preparation SHOULD be applied. However, the IRI resolver MUST also apply name preparation.

It is expected that the rules in Section 3.1 of [IRI] will be less restrictive than the rules for internationalized domain names, so that no escaping is necessary. Nevertheless, escaping is allowed for cases where not all characters can be directly represented.

4. Security Considerations

Besides the security considerations of [RFC 2396] and [IRI] and those applying to the various aspects of internationalized domain names in general, there are currently no known security problems.

Acknowledgements

To be done.

Copyright

Author's address

Martin J. Duerst
W3C/Keio University
5322 Endo, Fujisawa
252-8520 Japan
duerst@w3.org
http://www.w3.org/People/D%C3%BCrst/
Tel/Fax: +81 466 49 1170

Note: Please write "Duerst" with u-umlaut wherever possible, e.g. as "D&#252;rst" in XML and HTML.

References

[IRI] L. Masinter, M. Duerst, "Internationalized Resource Identifiers (IRI)", Internet Draft, January 2001, <http://www.ietf.org/internet-drafts/draft-masinter-url-i18n-06.txt>, work in progress.

[ISO10646] ISO/IEC, Information Technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane, Oct. 2000, with amendments.

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

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

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

[RFC 2277] H. Alvestrad, "IETF Policy on Character Sets and Languages".

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

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

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

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

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

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