Using the UTF-8 Character Set in the Domain Name System

Status of this Memo

This document is an Internet-Draft. 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."

To view the entire list of current Internet-Drafts, please check the "1id-abstracts.txt" listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa), ftp.nordu.net (Europe), munnari.oz.au (Pacific Rim), ds.internic.net (US East Coast), or ftp.isi.edu (US West Coast).

Abstract

The Domain Name System standard specifies that names are represented using the ASCII character encoding. This document expands that specification to allow the use of the UTF-8 character encoding, a superset of ASCII and a translation of the UCS-2 character encoding.
1. Introduction

The Domain Name System standard [RFC1035] specifies that names are represented using the ASCII character encoding. This document expands that specification to allow the use of the UTF-8 character encoding [RFC2044], a superset of ASCII and a translation of the UCS-2 character encoding.

Interpreting names as ASCII-only limits the utility of DNS in an international setting. The UTF-8 character set includes characters from most of the world’s written languages, allowing a far greater range of possible names and allowing names to use characters that are relevant to a particular locality. UTF-8 is the recommended character set for protocols that are evolving beyond ASCII [RFC2130].

This document defines the technology for a richer character set in DNS. It does not define the policy for the characters allowed in a name when used by a particular protocol. Protocol authors are encouraged to place no restrictions on characters allowed in a name.

2. Protocol Description

A UTF-8-aware DNS server is a DNS server that can load and store DNS names that contain UTF-8 characters. Names are encoded in logical order as opposed to visual order (see [UNICODE 2.0]).

Uniform downcasing permits UTF-8-aware DNS implementations to interoperate with non-UTF-8-aware DNS implementations. Any binary string can be used in a DNS name [RFC2181], but names must be compared with case-insensitivity [RFC1035]. A non-UTF-8-aware DNS implementation is unable to perform a case-insensitive comparison on a name containing UTF-8 characters. However, if UTF-8 names are downcased before transmission, then binary comparisons will provide the desired result on non-UTF-8-aware servers without violating the case-insensitivity requirement.

The DNS protocol standard states that original case should be preserved when possible as data is entered into the system. This requirement is modified as follows: a UTF-8-aware DNS server must downcase all names containing UTF-8 characters in both record names and record data before transmitting those names in any message. A UTF-8-aware DNS client/resolver must downcase all names containing UTF-8 characters before transmitting those names in any message.

For consistency, UTF-8-aware DNS servers must compare names that contain UTF-8 characters byte-for-byte, as opposed to using Unicode equivalency rules.
Applications should take care when allowing uppercase UTF-8 characters to be passed to the resolver, and DNS servers should take care when allowing uppercase UTF-8 characters to be entered in zone data. Downcasing in UTF-8 is locale-sensitive and the result may vary according to the locale of the code execution. The desired result will always be obtained if the application and server only accept lowercase characters.

Names encoded in UTF-8 must not exceed the size limits clarified in [RFC2181]: a maximum of 64 octets per label and 255 octets per name. Character count is insufficient to determine size, since some UTF-8 characters exceed one octet in length.

3. Interoperability Considerations

The UTF-8 character encoding is ideal for use with existing protocol implementations that expect US-ASCII characters. The representation of a US-ASCII character in UTF-8 is byte for byte identical to the US-ASCII representation. Non-UTF-8-aware DNS clients always encode names in ASCII format and those names will always be correctly interpreted by a UTF-8-aware DNS server.

DNS server authors may wish to provide a configuration switch on the DNS server to allow/disallow the use of UTF-8 characters on a per-server or per-zone basis.

A non-UTF-8-aware DNS server may accept a zone transfer of a zone containing UTF-8 names, but it may not be able to write back those names to a zone file or reload those names from a zone file. Administrators should exercise caution when transferring a zone containing UTF-8 names to a non-UTF-8-aware DNS server.

4. Security Considerations

The choice of character encoding for names does not impact the security of the DNS protocol.

5. Acknowledgements

The authors of this document would like to thank the following people for their contribution to this specification: John McConnell, Cliff Van Dyke and Bjorn Rettig.
6. References


7. Author’s Addresses

Stuart Kwan James Gilroy
Microsoft Corporation Microsoft Corporation
One Microsoft Way One Microsoft Way
Redmond, WA  98052 Redmond, WA  98052
USA  USA
<skwan@microsoft.com>  <jamesg@microsoft.com>