INTERNET-DRAFT

Stuart Kwan
James Gilroy
Levon Esibov
Microsoft Corp.

May 2001

<draft-skwan-utf8-dns-06.txt> Expires November 2001

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

Status of this Memo

Abstract

The Domain Names standard specifies that hostnames 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 Names standard [RFC1123] specifies that hostnames 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. This document specifically does not define policy for the characters allowed in a name when used in a particular application. For example, some protocols place restrictions on the characters allowed in a name

2. Protocol Description

2.1 Components and roles

Before the description of the protocol itself authors feel a need to clarify which components are involved in processing the hostnames and describe the usage of the hostnames by these components. The following list contains such information.

User.
User could be a human or application. Its role is to specify (also known as "write") and retrieve (also known as "read") the hostname to and from an application. The examples of such operations include typing the hostname, writing it on a touch sensitive screen, reading the name from the monitor, listening to a voicemail, etc...

Application.
Application's role is to
- process the hostname specified by user or other local or remote application.
- return to the user (for example display on a monitor screen) the hostname returned by DNS resolver.
- call DNS name resolution APIs to request resolver to perform the name resolution

Resolver.
Resolver's role is to
- process the name resolution requests from an application and submit appropriate DNS query to the DNS servers
- process the response from a DNS server and pass the response to the Application.

DNS server.
The role of the DNS server is to store and maintain the DNS data, process the updates to its database, update the replica copies of the databases and perform the DNS name resolution through responding to the DNS queries.

2.2 Protocol details

This section describes the modifications (if any) to each of these components and interfaces between the communicating components.

2.2.1 Users

No modifications to the users are proposed in this document. At the same time support of this protocol by other components specified later in this section may enable users to start using in hostnames characters from wider set than one specified in [RFC1123].

2.2.2 Interface between users and applications

User may use any character set or multiple character sets supported by the particular application. Specification of the allowed character sets supported by an application is outside of the scope of this document. The decision on which characters sets can be used to allow user to input and retrieve the hostnames is left to the implementers of the particular applications unless a protocol underlying specific application specifies the supported characters set. Thus this protocol does not affect the interface between users and applications.

2.2.3 Applications

Storage format of the hostnames by the applications is outside of the scope of this protocol.

2.2.4 Interface between applications and resolvers

This protocol does not specify the APIs that applications should use to request the resolver to perform the DNS name resolution of the internationalized hostnames. Instead it only specifies the format of the hostnames specified in the input and output of such APIs.

The applications supporting non-ASCII characters in hostnames MUST pass to the resolvers a hostname in ISO/IEC 10646 encoding. If the response returned by the resolver to the application contains the hostname, then the application should expect the hostname to be encoded using ISO/IEC 10646.

2.2.5 Resolvers

Before sending the hostname in the query packet, the resolver MUST prepare each name part as specified in [NAMEPREP]. After the name preparation the resolver MUST convert the hostname to be encoded using UTF-8 as specified in [RFC2044].
Names encoded in UTF-8 must not exceed the size limits clarified in [RFC2181]. Character count is insufficient to determine size, since some UTF-8 characters exceed one octet in length.

When resolver receives a response to the query from a DNS server, it MUST convert all of the hostnames from UTF-8 encoded format to the ISO/IEC 10646 encoding before passing these hostnames back to the application.

2.2.6 DNS servers

DNS servers authoritative for the records containing the hostnames containing the characters not allowed by [RFC1123] MUST allow use of the namepreped UTF-8 format to store and transmit those parts of the hostnames.

According to existing standards, any binary string can be used in a DNS name [RFC2181], but names must be compared with case-insensitivity [RFC1035]. At the same time DNS protocol standard states that original case SHOULD be preserved when possible as data is entered into the DNS database. This requirement is modified as follows: a DNS server authoritative for the internationalized hostnames MUST nameprep and perform UTF-8 conversion on all names containing internationalized characters in both record names and record data before storing these hostnames and transmitting those names in any message. This new requirement guarantees case-insensitive comparison of the internationalized hostnames even by those DNS servers that do not support this protocol.

DNS servers must compare names that contain UTF-8 characters byte-for-byte, as opposed to using Unicode equivalency rules.

3. Interoperability Considerations

If user continues using ASCII-only characters in the hostnames, then there is no need to upgrade any applications and/or resolvers.

As pointed in the previous section, there is no need to upgrade DNS servers, except possibly those that are authoritative for the zones containing internationalized hostnames.

The following interoperability issues should be taken into account

- A legacy application may not be able to process the hostnames containing non-ASCII characters returned by DNS resolvers. Effect of failure to process a name containing 7-bit needs to be separately investigated.
- If other protocols decide to use the nameprep-UTF-8-encoding to represent internationalized hostnames in their wire packets, then a legacy application supporting such protocol that receives UTF-8 encoded hostname from another application (for example, such as mail server or client) may fail to process such hostname. Effect of failure to process a name containing 7-bit needs to be separately investigate.

Thus hostnames that are intended to be globally usable [RFC1958] on legacy applications should still contain ASCII-only characters per [RFC1123].

- If an updated application runs on legacy resolver that rejects name resolution of the names containing any character not allowed by [RFC1123], then such resolvers will require an upgrade to enable name resolution of the internationalized hostnames.

- As specified above, DNS servers authoritative for the DNS records containing the internationalized hostnames must be able to save and load the hostnames containing napepreped-UTF-8-converted characters. If the DNS server doesn't satisfy this requirement, but needs to host such resource records, then it needs to be upgraded.

- Any DNS server involved in a name resolution process of the DNS records containing an internationalized hostname must not reject name resolution only because the hostname contains characters not allowed by [RFC1123]. This requirement does not mean that every DNS server in the name resolution path between the client and authoritative server must be able to store and load the DNS records containing the internationalized hostnames, but only means that the DNS server performing recursive resolution needs to be able to query for and cache such records, and that the DNS servers authoritative for the DNS names higher in the DNS name hierarchy than the internationalized names in query, need to be able to respond to such queries. Overwhelming majority of the DNS servers currently deployed on the Internet already satisfy this requirement. Authors are not aware of any implementation of the DNS server widely deployed on the Internet that doesn't satisfy this requirement.

Although most of the DNS servers may be capable of accepting a zone transfer of a zone containing UTF-8 encoded hostnames, some of them may not be able to store those names in a zone file or load those names from a zone file. Administrators should exercise caution when transferring a zone containing UTF-8 encoded hostnames to such DNS servers.

4. Security Considerations

Support for internationalized hostnames introduces a possibility of a new type of spoofing attacks that could be based on attacker's knowledge of misbehaving applications or resolvers that modifies the internationalized hostname that needs to be resolved. For example, if there is an application that modifies any character containing 7-bit in some predictable manner (for example by simply dropping the 7-bit),

then an attacker may register a DNS record mapping the derivative (i.e. modified by the misbehaving application or resolver) name to the data desired by attacker. In this scenario any user using such misbehaving application may receive as a result of name resolution the data (for example an IP address in A resource record) specified by the attacker without noticing that they are subjected to an attack even if the DNSSEC is used to verify the authenticity of the response.

Because this protocol depends on the procedures described in [NAMEPREP] and [RFC2044], the security issues identified in these document are also applicable to this 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

[RFC1035] P.V. Mockapetris, "Domain Names - Implementation and Specification," RFC 1035, ISI, Nov 1987.

[RFC2044] F. Yergeau, "UTF-8, a transformation format of Unicode and ISO 10646," RFC 2044, Alis Technologies, Oct 1996.

[RFC1958] B. Carpenter, "Architectural Principles of the Internet," RFC 1958, IAB, June 1996.

[RFC1123] R. Braden, "Requirements for Internet Hosts - Application and Support," STD 3, RFC 1123, January 1989.

[RFC2130] C. Weider et. al., "The Report of the IAB Character Set Workshop held 29 July - 1 March 1996", RFC 2130, Apr 1997.

[RFC2181] R. Elz and R. Bush, "Clarifications to the DNS Specification," RFC 2181, University of Melbourne and RGnet Inc, July 1997.

[UNICODE 2.0] The Unicode Consortium, "The Unicode Standard, Version 2.0," Addison-Wesley, 1996. ISBN 0-201-48345-9.

[NAMEPREP] Paul Hoffman and Marc Blanchet, "Preparation of Internationalized Host Names", draft-ietf-idn-nameprep-*.txt.

7. Author's Addresses

Stuart Kwan
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
USA
skwan@microsoft.com

James Gilroy
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
USA
jamesg@microsoft.com

Levon Esibov
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
USA
levone@microsoft.com

11. Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on the IETF's procedures with respect to rights in standards-track and standards-related documentation can be found in BCP-11. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementors or users of this specification can be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

12. Full Copyright Statement

except as needed for the purpose of developing Internet standards in