Network Working Group
Request for Comments: 3696
Category: Informational

J. Klensin

February 2004

Application Techniques for Checking and Transformation of Names

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Copyright Notice

Abstract

Many Internet applications have been designed to deduce top-level domains (or other domain name labels) from partial information. The introduction of new top-level domains, especially non-country-code ones, has exposed flaws in some of the methods used by these applications. These flaws make it more difficult, or impossible, for users of the applications to access the full Internet. This memo discusses some of the techniques that have been used and gives some guidance for minimizing their negative impact as the domain name environment evolves. This document draws summaries of the applicable rules together in one place and supplies references to the actual standards.

Table of Contents

1. Introduction

Designers of user interfaces to Internet applications have often found it useful to examine user-provided values for validity before passing them to the Internet tools themselves. This type of test, most commonly involving syntax checks or application of other rules to domain names, email addresses, or "web addresses" (URLs or, occasionally, extended URI forms (see Section 4)) may enable better-quality diagnostics for the user than might be available from the protocol itself. Local validity tests on values are also thought to improve the efficiency of back-office processing programs and to reduce the load on the protocols themselves. Certainly, they are consistent with the well-established principle that it is better to detect errors as early as possible.

The tests must, however, be made correctly or at least safely. If criteria are applied that do not match the protocols, users will be inconvenienced, addresses and sites will effectively become inaccessible to some groups, and business and communications opportunities will be lost. Experience in recent years indicates that syntax tests are often performed incorrectly and that tests for top-level domain names are applied using obsolete lists and conventions. We assume that most of these incorrect tests are the result of the inability to conveniently locate exact definitions for the criteria to be applied. This document draws summaries of the applicable rules together in one place and supplies references to the

actual standards. It does not add anything to those standards; it merely draws the information together into a form that may be more accessible.

Many experts on Internet protocols believe that tests and rules of these sorts should be avoided in applications and that the tests in the protocols and back-office systems should be relied on instead. Certainly implementations of the protocols cannot assume that the data passed to them will be valid. Unless the standards specify particular behavior, this document takes no position on whether or not the testing is desirable. It only identifies the correct tests to be made if tests are to be applied.

The sections that follow discuss domain names, email addresses, and URLs.

2. Restrictions on domain (DNS) names

The authoritative definitions of the format and syntax of domain names appear in RFCs 1035 [RFC1035], 1123 [RFC1123], and 2181 [RFC2181].

Any characters, or combination of bits (as octets), are permitted in DNS names. However, there is a preferred form that is required by most applications. This preferred form has been the only one permitted in the names of top-level domains, or TLDs. In general, it is also the only form permitted in most second-level names registered in TLDs, although some names that are normally not seen by users obey other rules. It derives from the original ARPANET rules for the naming of hosts (i.e., the "hostname" rule) and is perhaps better described as the "LDH rule", after the characters that it permits. The LDH rule, as updated, provides that the labels (words or strings separated by periods) that make up a domain name must consist of only the ASCII [ASCII] alphabetic and numeric characters, plus the hyphen. No other symbols or punctuation characters are permitted, nor is blank space. If the hyphen is used, it is not permitted to appear at either the beginning or end of a label. There is an additional rule that essentially requires that top-level domain names not be all-numeric.

When it is necessary to express labels with non-character octets, or to embed periods within labels, there is a mechanism for keying them in that utilizes an escape sequence. RFC 1035 [RFC1035] should be consulted if that mechanism is needed (most common applications, including email and the Web, will generally not permit those escaped strings). A special encoding is now available for non-ASCII characters, see the brief discussion in Section 5.

Most internet applications that reference other hosts or systems assume they will be supplied with "fully-qualified" domain names, i.e., ones that include all of the labels leading to the root, including the TLD name. Those fully-qualified domain names are then passed to either the domain name resolution protocol itself or to the remote systems. Consequently, purported DNS names to be used in applications and to locate resources generally must contain at least one period (".") character. Those that do not are either invalid or require the application to supply additional information. Of course, this principle does not apply when the purpose of the application is to process or query TLD names themselves. The DNS specification also permits a trailing period to be used to denote the root, e.g., "a.b.c" and "a.b.c." are equivalent, but the latter is more explicit and is required to be accepted by applications. This convention is especially important when a TLD name is being referred to directly. For example, while ".COM" has become the popular terminology for referring to that top-level domain, "COM." would be strictly and technically correct in talking about the DNS, since it shows that "COM" is a top-level domain name.

There is a long history of applications moving beyond the "one or more periods" test in an attempt to verify that a valid TLD name is actually present. They have done this either by applying some heuristics to the form of the name or by consulting a local list of valid names. The historical heuristics are no longer effective. If one is to keep a local list, much more effort must be devoted to keeping it up-to-date than was the case several years ago.

The heuristics were based on the observation that, since the DNS was first deployed, all top-level domain names were two, three, or four characters in length. All two-character names were associated with "country code" domains, with the specific labels (with a few early exceptions) drawn from the ISO list of codes for countries and similar entities [IS3166]. The three-letter names were "generic" TLDs, whose function was not country-specific, and there was exactly one four-letter TLD, the infrastructure domain "ARPA." [RFC1591]. However, these length-dependent rules were conventions, rather than anything on which the protocols depended.

Before the mid-1990s, lists of valid top-level domain names changed infrequently. New country codes were gradually, and then more rapidly, added as the Internet expanded, but the list of generic domains did not change at all between the establishment of the "INT." domain in 1988 and ICANN's allocation of new generic TLDs in 2000. Some application developers responded by assuming that any two-letter domain name could be valid as a TLD, but the list of generic TLDs was fixed and could be kept locally and tested. Several of these assumptions changed as ICANN started to allocate new top-level

domains: one two-letter domain that does not appear in the ISO 3166-1 table [ISO.3166.1988] was tentatively approved, and new domains were created with three, four, and even six letter codes.

As of the first quarter of 2003, the list of valid, non-country, top-level domains was .AERO, .BIZ, .COM, .COOP, .EDU, .GOV, .INFO, .INT, .MIL, .MUSEUM, .NAME, .NET, .ORG, .PRO, and .ARPA. ICANN is expected to expand that list at regular intervals, so the list that appears here should not be used in testing. Instead, systems that filter by testing top-level domain names should regularly update their local tables of TLDs (both "generic" and country-code-related) by polling the list published by IANA [DomainList]. It is likely that the better strategy has now become to make the "at least one period" test, to verify LDH conformance (including verification that the apparent TLD name is not all-numeric), and then to use the DNS to determine domain name validity, rather than trying to maintain a local list of valid TLD names.

A DNS label may be no more than 63 octets long. This is in the form actually stored; if a non-ASCII label is converted to encoded "punycode" form (see Section 5), the length of that form may restrict the number of actual characters (in the original character set) that can be accommodated. A complete, fully-qualified, domain name must not exceed 255 octets.

Some additional mechanisms for guessing correct domain names when incomplete information is provided have been developed for use with the web and are discussed in Section 4.4.

3. Restrictions on email addresses

Reference documents: RFC 2821 [RFC2821] and RFC 2822 [RFC2822]

Contemporary email addresses consist of a "local part" separated from a "domain part" (a fully-qualified domain name) by an at-sign ("@"). The syntax of the domain part corresponds to that in the previous section. The concerns identified in that section about filtering and lists of names apply to the domain names used in an email context as well. The domain name can also be replaced by an IP address in square brackets, but that form is strongly discouraged except for testing and troubleshooting purposes.

The local part may appear using the quoting conventions described below. The quoted forms are rarely used in practice, but are required for some legitimate purposes. Hence, they should not be rejected in filtering routines but, should instead be passed to the email system for evaluation by the destination host.

The exact rule is that any ASCII character, including control characters, may appear quoted, or in a quoted string. When quoting is needed, the backslash character is used to quote the following character. For example

Abc\@def@example.com

is a valid form of an email address. Blank spaces may also appear, as in

Fred\ Bloggs@example.com

The backslash character may also be used to quote itself, e.g.,

Joe.\\Blow@example.com

In addition to quoting using the backslash character, conventional double-quote characters may be used to surround strings. For example

"Abc@def"@example.com

"Fred Bloggs"@example.com

are alternate forms of the first two examples above. These quoted forms are rarely recommended, and are uncommon in practice, but, as discussed above, must be supported by applications that are processing email addresses. In particular, the quoted forms often appear in the context of addresses associated with transitions from other systems and contexts; those transitional requirements do still arise and, since a system that accepts a user-provided email address cannot "know" whether that address is associated with a legacy system, the address forms must be accepted and passed into the email environment.

Without quotes, local-parts may consist of any combination of alphabetic characters, digits, or any of the special characters

! # $ % & ' * + - / = ? ^ _ ` . { | } ~

period (".") may also appear, but may not be used to start or end the local part, nor may two or more consecutive periods appear. Stated differently, any ASCII graphic (printing) character other than the at-sign ("@"), backslash, double quote, comma, or square brackets may appear without quoting. If any of that list of excluded characters are to appear, they must be quoted. Forms such as

user+mailbox@example.com

customer/department=shipping@example.com

$A12345@example.com

!def!xyz%abc@example.com

_somename@example.com

are valid and are seen fairly regularly, but any of the characters listed above are permitted. In the context of local parts, apostrophe ("'") and acute accent ("`") are ordinary characters, not quoting characters. Some of the characters listed above are used in conventions about routing or other types of special handling by some receiving hosts. But, since there is no way to know whether the remote host is using those conventions or just treating these characters as normal text, sending programs (and programs evaluating address validity) must simply accept the strings and pass them on.

In addition to restrictions on syntax, there is a length limit on email addresses. That limit is a maximum of 64 characters (octets) in the "local part" (before the "@") and a maximum of 255 characters (octets) in the domain part (after the "@") for a total length of 320 characters. Systems that handle email should be prepared to process addresses which are that long, even though they are rarely encountered.

4. URLs and URIs

4.1. URI syntax definitions and issues

The syntax for URLs (Uniform Resource Locators) is specified in [RFC1738]. The syntax for the more general "URI" (Uniform Resource Identifier) is specified in [RFC2396]. The URI syntax is extremely general, with considerable variations permitted according to the type of "scheme" (e.g., "http", "ftp", "mailto") that is being used. While it is possible to use the general syntax rules of RFC 2396 to perform syntax checks, they are general enough --essentially only specifying the separation of the scheme name and "scheme specific part" with a colon (":") and excluding some characters that must be escaped if used-- to provide little significant filtering or validation power.

The following characters are reserved in many URIs -- they must be used for either their URI-intended purpose or must be encoded. Some particular schemes may either broaden or relax these restrictions (see the following sections for URLs applicable to "web pages" and electronic mail), or apply them only to particular URI component parts.

; / ? : @ & = + $, ?

In addition, control characters, the space character, the double-quote (") character, and the following special characters

< > # %

are generally forbidden and must either be avoided or escaped, as discussed below.

The colon after the scheme name, and the percent sign used to escape characters, are specifically reserved for those purposes, although ":" may also be used elsewhere in some schemes.

When it is necessary to encode these, or other, characters, the method used is to replace it with a percent-sign ("%") followed by two hexidecimal digits representing its octet value. See section 2.4.1 of [RFC2396] for an exact definition. Unless it is used as a delimiter of the URI scheme itself, any character may optionally be encoded this way; systems that are testing URI syntax should be prepared for these encodings to appear in any component of the URI except the scheme name itself.

A "generic URI" syntax is specified and is more restrictive, but using it to test URI strings requires that one know whether or not the particular scheme in use obeys that syntax. Consequently, applications that intend to check or validate URIs should normally identify the scheme name and then apply scheme-specific tests. The rules for two of those -- HTTP [RFC1738] and MAILTO [RFC2368] URLs -- are discussed below, but the author of an application which intends to make very precise checks, or to reject particular syntax rather than just warning the user, should consult the relevant scheme-definition documents for precise syntax and relationships.

4.2. The HTTP URL

Absolute HTTP URLs consist of the scheme name, a host name (expressed as a domain name or IP address), and optional port number, and then, optionally, a path, a search part, and a fragment identifier. These are separated, respectively, by a colon and the two slashes that precede the host name, a colon, a slash, a question mark, and a hash mark ("#"). So we have

http://host:port/path?search#fragment

http://host/path/

http://host/path#fragment

http://host/path?search

http://host

and other variations on that form. There is also a "relative" form, but it almost never appears in text that a user might, e.g., enter into a form. See [RFC2616] for details.

The characters

/ ; ?

are reserved within the path and search parts and must be encoded; the first of these may be used unencoded, and is often used within the path, to designate hierarchy.

4.3. The MAILTO URL

MAILTO is a URL type whose content is an email address. It can be used to encode any of the email address formats discussed in Section 3 above. It can also support multiple addresses and the inclusion of headers (e.g., Subject lines) within the body of the URL. MAILTO is authoritatively defined in RFC 2368 [RFC2368]; anyone expecting to accept and test multiple addresses or mail header or body formats should consult that document carefully.

In accepting text for, or validating, a MAILTO URL, it is important to note that, while it can be used to encode any valid email address, it is not sufficient to copy an email address into a MAILTO URL since email addresses may include a number of characters that are invalid in, or have reserved uses for, URLs. Those characters must be encoded, as outlined in Section 4.1 above, when the addresses are mapped into the URL form. Conversely, addresses in MAILTO URLs cannot, in general, be copied directly into email contexts, since few email programs will reverse the decodings (and doing so might be interpreted as a protocol violation).

The following characters may appear in MAILTO URLs only with the specific defined meanings given. If they appear in an email address (i.e., for some other purpose), they must be encoded:

: The colon in "mailto:"

< > # " % { } | \ ^ ~ `

These characters are "unsafe" in any URL, and must always be encoded.

The following characters must also be encoded if they appear in a MAILTO URL

? & =
Used to delimit headers and their values when these are encoded into URLs.

Some examples may be helpful:

Email address	MAILTO URL	Notes
Joe@example.com	mailto:joe@example.com	1
user+mailbox@example.com	mailto:user%2Bmailbox@example.com	2
customer/department=shipping@example.com	mailto:customer%2Fdepartment=shipping@example.com	3
$A12345@example.com	mailto:$A12345@example.com	4
!def!xyz%abc@example.com	mailto:!def!xyz%25abc@example.com	5
_somename@example.com	mailto:_somename@example.com	4

Table 1

Notes on Table

1. No characters appear in the email address that require escaping, so the body of the MAILTO URL is identical to the email address.

2. There is actually some uncertainty as to whether or not the "+" characters requires escaping in MAILTO URLs (the standards are not precisely clear). But, since any character in the address specification may optionally be encoded, it is probably safer to encode it.

3. The "/" character is generally reserved in URLs, and must be encoded as %2F.

4. Neither the "$" nor the "_" character are given any special interpretation in MAILTO URLs, so need not be encoded.

5. While the "!" character has no special interpretation, the "%" character is used to introduce encoded sequences and hence it must always be encoded.

4.4. Guessing domain names in web contexts

Several web browsers have adopted a practice that permits an incomplete domain name to be used as input instead of a complete URL. This has, for example, permitted users to type "microsoft" and have the browser interpret the input as "http://www.microsoft.com/". Other browser versions have gone even further, trying to build DNS names up through a series of heuristics, testing each variation in turn to see if it appears in the DNS, and accepting the first one found as the intended domain name. Still, others automatically invoke search engines if no period appears or if the reference fails. If any of these approaches are to be used, it is often critical that the browser recognize the complete list of TLDs. If an incomplete list is used, complete domain names may not be recognized as such and the system may try to turn them into completely different names. For example, "example.aero" is a fully-qualified name, since "AERO." is a TLD name. But, if the system doesn't recognize "AERO" as a TLD name, it is likely to try to look up "example.aero.com" and "www.example.aero.com" (and then fail or find the wrong host), rather than simply looking up the user-supplied name.

As discussed in Section 2 above, there are dangers associated with software that attempts to "know" the list of top-level domain names locally and take advantage of that knowledge. These name-guessing heuristics are another example of that situation: if the lists are up-to-date and used carefully, the systems in which they are embedded may provide an easier, and more attractive, experience for at least some users. But finding the wrong host, or being unable to find a host even when its name is precisely known, constitute bad experiences by any measure.

More generally, there have been bad experiences with attempts to "complete" domain names by adding additional information to them. These issues are described in some detail in RFC 1535 [RFC1535].

5. Implications of internationalization

The IETF has adopted a series of proposals ([RFC3490] - [RFC3492]) whose purpose is to permit encoding internationalized (i.e., non-ASCII) names in the DNS. The primary standard, and the group generically, are known as "IDNA". The actual strings stored in the

DNS are in an encoded form: the labels begin with the characters "xn--" followed by the encoded string. Applications should be prepared to accept and process the encoded form (those strings are consistent with the "LDH rule" (see Section 2) so should not raise any separate issues) and the use of local, and potentially other, characters as appropriate to local systems and circumstances.

The IDNA specification describes the exact process to be used to validate a name or encoded string. The process is sufficiently complex that shortcuts or heuristics, especially for versions of labels written directly in Unicode or other coded character sets, are likely to fail and cause problems. In particular, the strings cannot be validated with syntax or semantic rules of any of the usual sorts: syntax validity is defined only in terms of the result of executing a particular function.

In addition to the restrictions imposed by the protocols themselves, many domains are implementing rules about just which non-ASCII names they will permit to be registered (see, e.g., [JET], [RegRestr]). This work is still relatively new, and the rules and conventions are likely to be different for each domain, or at least each language or script group. Attempting to test for those rules in a client program to see if a user-supplied name might possibly exist in the relevant domain would almost certainly be ill-advised.

One quick local test however, may be reasonable: as of the time of this writing, there should be no instances of labels in the DNS that start with two characters, followed by two hyphens, where the two characters are not "xn" (in, of course, either upper or lower case). Such label strings, if they appear, are probably erroneous or obsolete, and it may be reasonable to at least warn the user about them.

There is ongoing work in the IETF and elsewhere to define internationalized formats for use in other protocols, including email addresses. Those forms may or may not conform to existing rules for ASCII-only identifiers; anyone designing evaluators or filters should watch that work closely.

6. Summary

When an application accepts a string from the user and ultimately passes it on to an API for a protocol, the desirability of testing or filtering the text in any way not required by the protocol itself is hotly debated. If it must divide the string into its components, or otherwise interpret it, it obviously must make at least enough tests to validate that process. With, e.g., domain names or email addresses that can be passed on untouched, the appropriateness of

trying to figure out which ones are valid and which ones are not requires a more complex decision, one that should include considerations of how to make exactly the correct tests and to keep information that changes and evolves up-to-date. A test containing obsolete information, can be extremely frustrating for potential correspondents or customers and may harm desired relationships.

7. Security Considerations

Since this document merely summarizes the requirements of existing standards, it does not introduce any new security issues. However, many of the techniques that motivate the document raise important security concerns of their own. Rejecting valid forms of domain names, email addresses, or URIs often denies service to the user of those entities. Worse, guessing at the user's intent when an incomplete address, or other string, is given can result in compromises to privacy or accuracy of reference if the wrong target is found and returned. From a security standpoint, the optimum behavior is probably to never guess, but instead, to force the user to specify exactly what is wanted. When that position involves a tradeoff with an acceptable user experience, good judgment should be used and the fact that it is a tradeoff recognized.

Some characters have special or privileged meanings on some systems (i.e., ` on Unix). Applications should be careful to escape those locally if necessary. By the same token, they are valid, and should not be disallowed locally, or escaped when transmitted through Internet protocols, for such reasons if a remote site chooses to use them.

The presence of local checking does not permit remote checking to be bypassed. Note that this can apply to a single machine; in particular, a local MTA should not assume that a local MUA has properly escaped locally-significant special characters.

8. Acknowledgements

The author would like to express his appreciation for helpful comments from Harald Alvestrand, Eric A. Hall, and the RFC Editor, and for partial support of this work from SITA. Responsibility for any errors remains, of course, with the author.

The first Internet-Draft on this subject was posted in February 2003. The document was submitted to the RFC Editor on 20 June 2003, returned for revisions on 19 August, and resubmitted on 5 September 2003.

9. References

9.1. Normative References

[RFC1035] Mockapetris, P., "Domain names - implementation and specification", STD 13, RFC 1035, November 1987.

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

[RFC1535] Gavron, E., "A Security Problem and Proposed Correction With Widely Deployed DNS Software", RFC 1535, October 1993.

[RFC1738] Berners-Lee, T., Masinter, L. and M. McCahill, "Uniform Resource Locators (URL)", RFC 1738, December 1994.

[RFC2181] Elz, R. and R. Bush, "Clarifications to the DNS Specification", RFC 2181, July 1997.

[RFC2368] Hoffman, P., Masinter, L. and J. Zawinski, "The mailto URL scheme", RFC 2368, July 1998.

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

[RFC2616] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", RFC 2616, June 1999.

[RFC2821] Klensin, J., Ed., "Simple Mail Transfer Protocol", RFC 2821, April 2001.

[RFC2822] Resnick, P., Ed., "Internet Message Format", RFC 2822, April 2001.

[RFC3490] Faltstrom, P., Hoffman, P. and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, March 2003.

[RFC3491] Hoffman, P. and M. Blanchet, "Nameprep: A Stringprep Profile for Internationalized Domain Names (IDN)", RFC 3491, March 2003.

[RFC3492] Costello, A., "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)", RFC 3492, March 2003.

[ASCII] American National Standards Institute (formerly United States of America Standards Institute), "USA Code for Information Interchange", ANSI X3.4-1968. ANSI X3.4-1968 has been replaced by newer versions with slight modifications, but the 1968 version remains definitive for the Internet.

[DomainList] Internet Assigned Numbers Authority (IANA), Untitled alphabetical list of current top-level domains. http://data.iana.org/TLD/tlds-alpha-by-domain.txt ftp://data.iana.org/TLD/tlds-alpha-by-domain.txt

9.2. Informative References

[ISO.3166.1988] International Organization for Standardization, "Codes for the representation of names of countries, 3rd edition", ISO Standard 3166, August 1988.

[JET] Konishi, K., et al., "Internationalized Domain Names Registration and Administration Guideline for Chinese, Japanese and Korean", Work in Progress.

[RFC1591] Postel, J., "Domain Name System Structure and Delegation", RFC 1591, March 1994.

[RegRestr] Klensin, J., "Registration of Internationalized Domain Names: Overview and Method", Work in Progress, February 2004.

10. Author's Address

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

Phone: +1 617 491 5735
EMail: john-ietf@jck.com

11. Full Copyright Statement

Intellectual Property

Acknowledgement

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