Abstract

Internationalized Domain Names for Applications (IDNA) provides a method to map a subset of names written in Unicode into the DNS. Because of Unicode decisions, appearance, language and writing system conventions, and historical reasons, it often has been asserted that there is more than one way to write what competent readers and writers think of as the same host name; these different ways of writing are often called "variants". (The authors note that there are many conflicting definitions for the term "variant" in the IDNA community.) This document surveys the approaches that top-level domains have taken to the registration and provisioning of domain names that have variants. This document is not a product of the IETF, does not propose any method to make variants work "correctly", and is not an introduction to internationalization or IDNA.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This is a contribution to the RFC Series, independently of any other RFC stream. The RFC Editor has chosen to publish this document at its discretion and makes no statement about its value for implementation or deployment. Documents approved for publication by the RFC Editor are not a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6927.
Table of Contents

1. Introduction .................................................... 3
2. Terminology ..................................................... 4
3. Base Documents ................................................ 5
4. Domain Practices of gTLDs ..................................... 6
   4.1. AERO .................................................... 6
   4.2. ASIA ..................................................... 6
   4.3. BIZ ...................................................... 6
   4.4. CAT ...................................................... 6
   4.5. COM ...................................................... 7
   4.6. COOP ..................................................... 7
   4.7. INFO ..................................................... 7
   4.8. JOBS ..................................................... 7
   4.9. MOBI ..................................................... 7
   4.10. MUSEUM ................................................. 8
   4.11. NAME ................................................... 8
   4.12. NET ..................................................... 8
   4.13. ORG ..................................................... 8
   4.14. POST ................................................... 9
   4.15. PRO ..................................................... 9
   4.16. TEL ................................................... 9
   4.17. TRAVEL ................................................ 10
   4.18. XXX ................................................... 10
5. Domain Practices of ccTLDs ................................... 10
   5.1. BG ....................................................... 10
   5.2. BR ....................................................... 10
   5.3. CL ....................................................... 10
   5.4. CN ....................................................... 10
   5.5. ES ....................................................... 11
   5.6. EU ....................................................... 11
   5.7. GR ....................................................... 11
   5.8. IL ....................................................... 11
   5.9. IR ....................................................... 11
   5.10. JP ...................................................... 11
   5.11. KR ...................................................... 12
1. Introduction

Internationalized Domain Names for Applications (IDNA) [RFC5890] allows host names in the DNS [RFC1035] to contain characters from the Unicode repertoire. Some Unicode characters are considered to be "variants" of one another. Because of the 20th century reform of Chinese writing, there is often more than one representation of what Chinese speakers think of as the same character. Some languages written in Latin characters with accents and diacritical marks, known as decorated characters, allow the decorations to be omitted in some situations; for example, French sometimes omits accents on capital letters, depending on country and culture. Due to the difficulty of representing decorated characters in ASCII systems, many users have informally used undecorated characters in DNS host names, even when they are not linguistically equivalent to the decorated versions.

There is no single agreed-on definition of "variant". In 2012, ICANN said that variants "occur when a single conceptual character can be identified with two or more different Unicode Code Points with graphic representations that may be visually similar" (this definition was previously available at http://www.icann.org/en/resources/idn/variant-tlds). ICANN’s IDN Variant Issues Project report [VIPREPORT] says that "[t]here is today no fully accepted definition for what may constitute a variant relationship between top-level labels". RFC 3743 [RFC3743] (an Informational RFC, not the product of the IETF) says that the idea of variants is "wherein one conceptual character can be identified with several different Code Points in character sets for computer use".

The proper handing of variant names has been a topic of extensive debate and research, with little consensus reached on how to handle them or even what characters are variants of each other. Many people...
would like variant names to behave "the same", for a diverse range of meanings of "same". In some cases, it is a textual similarity, such as variants having corresponding DNS records; in some, it is functional similarity, such as variant names resolving to the same web server; while in others, it is user experience similarity, such as names resolving to web sites that, while not identical, are perceived by human users as equivalent.

This document provides a snapshot of variant handling in the top-level domains (TLDs) contracted by ICANN, so-called gTLDs (generic TLDs) and sTLDs (sponsored TLDs), as of late 2012. We chose those domains because ICANN requires each TLD to describe its IDN and variant practices, and the TLD zone files are available for inspection, to verify what actually goes into the zones. This document also contains a small sampling of so-called ccTLDs (country code TLDs, the TLDs that consist of two ASCII letters) for which we could find information.

Since "variant" can mean vastly different things to different people, there is also no agreement about when two zones are supposed to "behave the same". Also, the gTLDs and sTLDs might have different views of what variants are and are not required to report to ICANN about their policies.

2. Terminology

We use some terminology that has become generally agreed to when discussing variant names, although we openly admit that such agreement is not complete and the terminology continues to change.

Bundle: The IDN practices documents (see below) can identify sets of code points that are considered variants of each other using Language Variant Tables, defined in [RFC3743]. A set of names in which the characters in each position are variants is known as a bundle or, more technically, as an "IDL Package". The variant rules vary among languages, and for the same language can vary among TLDs. Many languages do not define variant characters and hence do not have bundles.

Allocated: A name is allocated if sponsorship of that label in some zone has been granted. This is similar to what many people refer to as "registered".

Active: A name is active if it appears as an owner name in a zone. Most allocated names are active, but some are not.
Blocked: Some names cannot be registered at all. For example, some registries allow one name in a bundle to be registered and block the rest.

Withheld: Some names can only be allocated under certain conditions. For example, some registries permit only the registrant of one name in a bundle to register or activate other names in the same bundle.

Parallel NS: Multiple names in a bundle are provisioned in the TLD with identical NS records, so they all are handled by the same name servers.

DNAME aliasing: The DNAME [RFC6672] DNS record creates a shadow tree of DNS records, roughly as though there were a CNAME in the shadow tree pointing to each name in the target tree. DNAMEs have been used both to provide resolution for several names in a bundle and to provide resolution for every name under a TLD.

3. Base Documents

ICANN has published a variety of documents on variant management. The most important are the "Guidelines for the Implementation of Internationalized Domain Names" issued in Version 1.0 [G1] and Version 3.0 [G3].

ICANN says that TLDs are supposed to register an IDN practices document with IANA for each language and/or script in which the TLD accepts IDN registrations, to be entered in the IANA Repository of IDN Practices [IANAIDN]. The practices document lists the Unicode characters allowed in names in the language or script, which characters are considered equivalent, and which of an equivalent group is preferred. Some TLDs have been more diligent than others at keeping the registry up to date. Also, some TLDs have tables for a few languages and scripts, while others (notably .COM, .NET, and .NAME) have a large set of tables, including some for languages and scripts that are no longer spoken or used, such as Runic and Ogham. The authors also note that many of the tables in the IANA registry are clearly out of date, containing URLs of policy pages that no longer exist and contact information for people who have left the registry.

Some of the ICANN agreements with each TLD [ICANNAGREE] describe the TLD’s IDN practices, but most don’t.
4. Domain Practices of gTLDs

This list covers most of the current set of gTLDs. In most cases, the authors have also checked the zone files for the gTLD to verify or augment the policy description.

4.1. AERO

The .AERO TLD has no IDNs and no rules or practices for them.

4.2. ASIA

The .ASIA domain accepts registrations in many Asian languages. They have IANA tables for Japanese, Korean, and Chinese. The IANA tables refer to their CJK IDN policies [ASIACJK], which say that applied-for and preferred IDN variants are "active and included in the zone". No IDN publication mechanism is described in the documentation, but since the zone file contains no DNAMEs, they must be using parallel NS for variants.

4.3. BIZ

ICANN gave the registry (Neustar) non-specific permission to register IDNs in a letter in 2004 [TWOMEY04A]. The IDN rules were apparently discussed with ICANN but not defined; see Appendix 9 of the registry agreement [ICANNBIZ9].

They have about a dozen IANA tables. No IDN publication mechanism is described, but from inspection, it appears that variants are blocked.

4.4. CAT

The IDN rules are described in Appendix S, Part VII.2 [ICANNCATS] of the ICANN agreement. "Registry will take a very cautious approach in its IDN offerings. IDNs will be bundled with the equivalent ASCII domains". The only language is Catalan. No IDN publication mechanism is described.

Appendix S includes "The list of non-ASCII-characters for Catalan language and their ASCII equivalent for the purposes of the defined service", which implicitly describes bundles. The bundles consist of names with accented and unaccented vowels, U+00E7 ("c with cedilla") and a plain c, and the Catalan "ela geminada" written as two l's separated by a U+00B7 ("middle dot") and the three characters "l-l".

Levine & Hoffman              Informational                     [Page 6]
When a registrant registers an IDN, the registry also includes the ASCII version. From inspection of the zone file, the ASCII version is provisioned with NS, and the IDN is a DNAME alias of the ASCII version.

4.5. COM

ICANN and Verisign have extensive correspondence about IDNs and variants, including letters to ICANN from Ben Turner [TURNER03] and Russell Lewis [LEWIS03].

The IANA registry has tables for several dozen languages, including archaic languages such as hieroglyphics and Aramaic. Verisign publishes documents describing scripts and languages [VRSNLANG], character variants [VRSNCHAR], registration rules [VRSNRULES], and additional registration logic [VRSNADDL].

In Chinese, variants are blocked (see [VRSNADDL]). In other languages, there is no bundling or blocking.

4.6. COOP

The .COOP TLD has no IDNs and no rules or practices for them.

4.7. INFO

The IANA registry has a table for German. The German table notes that "the Eszet ... character used in the German script will be mapped to a double ‘s’ string (i.e. ‘ss’)"). The domain also offers names in Greek, Russian, Arabic, Korean, and other languages. The list and IDN tables are on the registry’s web site [INFOTABLES].

Afilias says (not in a published policy) that it does not allow Korean characters with different widths and that there are no variants in .INFO.

Appendix 9 of the registry agreement [ICANNINFO9] refers to a 2003 letter from Paul Twomey [TWOMEY03] that refers to blocking variants.

4.8. JOBS

The .JOBS TLD has no IDNs and no rules or practices for them.

4.9. MOBI

The zone file has about 22,000 IDNs. Afilias says (not in a published policy) that .MOBI supports Simplified Chinese only and that the language table for this is the same as that used by .CN.
Variant characters are blocked from registration. The domain has no tables at IANA. Appendix S of the registry agreement [ICANNMOBIS], says that IDNs are provisioned according to [G1].

4.10. MUSEUM

The zone file has many IDNs, but spot checks find that many are lame or dead. A 2004 letter from Paul Twomey [TWOMEY04] refers to [G1].

The registry has a detailed policy page [MUSEUMPOLICY]. IDNs are accepted in Latin and Hebrew scripts, with plans for Arabic, Chinese, Japanese, Korean, Cyrillic, and Greek. They do no bundling or blocking, but names that may be confusable due to visual similarity are not allowed. These are apparently determined by manual inspection, which is practical due to the very small size of the domain.

4.11. NAME

The .NAME TLD is managed the same as .COM.

4.12. NET

The .NET TLD is managed the same as .COM.

4.13. ORG

A 2003 letter from Paul Twomey [TWOMEY03A] refers to [G1]. The registry has a list of IDN languages [PIRIDN], several written in Latin script, plus Chinese and Korean. A Questions page [PIRFAQ] states that Chinese names have been accepted since January 2010 and Cyrillic names in seven languages since February 2011. The practices for some, but not all, of the Latin languages are registered with IANA.

A Chinese language policy form on the Public Interest Registry (PIR) web site says that the ZH-CN and ZH-TW IDNs use the corresponding ccTLD tables from IANA, and check boxes say that Variant Registration Policies and Variant Management Policies are applicable but don’t say what those policies are.

Private correspondence [CHANDIWALA12] describes not-yet-public rules for variants in Chinese and Cyrillic in .ORG that restrict the number of variants that a registration can have.

The Korean language policy form says that it uses the KRNIC table for Korean from IANA and that there are no variants.
4.14. POST

The .POST TLD appears to have no registrations at all yet.

4.15. PRO

The .PRO TLD has no IDNs and no rules or practices for them.

4.16. TEL

The zone has many IDNs. It is probably operating according to a 2004 letter from Paul Twomey [TWOMEY04A] to Neustar, which did not mention specific TLDs. Its policy page [TELPOLICY] has links to IDN practices for 17 languages, all but one of which are registered with IANA. None of the Latin scripts do bundling or blocking. The Japanese practices say that variants are blocked. The Chinese practices document says:

Therefore, in addition to the blocking mechanism, bundling is also implemented for the Chinese language IDNs. When registering a Chinese language IDN (primary domain name) up to two additional variant domain names will be automatically registered. The first variant will consist entirely of simplified Chinese characters that correspond to those comprising the primary domain name. The second variant will consist exclusively of traditional Chinese characters that correspond to those comprising the primary domain name.

The primary domain name together with the requested variants constitutes a bundle on which all operations are atomic. For example, if the registrant adds a name server to the primary domain name, all names in the bundle will be associated with that new name server.

The zone has no DNAME records, so the second paragraph strongly suggests parallel NS.

The .TEL TLD, intended as an online directory, does not allow registrants to enter arbitrary Resource Records (RRs) in the zone. Nearly all names have NS records pointing to Telnic’s own name servers. The A records all point to Telnic’s own web server that shows directory information. NAPTR records provide telephone numbers of registrants if available. Users can only directly provision MX records. Currently, there are 16 domains, none of which are IDNs, that point to random other name servers and mostly appear to be parked.
4.17. TRAVEL

The .TRAVEL TLD has no IDNs and no rules or practices for them.

4.18. XXX

The .XXX TLD has no IDNs and no rules or practices for them.

5. Domain Practices of ccTLDs

Some ccTLDs publish their IDN policies. This section is a non-exhaustive sampling of some of those policies. Note that few ccTLDs make their zone files available, so the authors could not validate the policies by looking in the zone files.

5.1. BG

The .BG TLD (for Bulgaria) publishes a policy page [BGPOLICY]. It has published an IDN table for the Bulgarian and Russian languages in [IANAIDN]. The policy does not mention variants.

5.2. BR

The .BR TLD (for Brazil) publishes a policy page [BRPOLICY]. It has published an IDN table for the Portuguese language in [IANAIDN]. Although the IDN table does not describe variants, the policy page says that bundles consist of names that are the same disregarding accents on vowels, cedillas on letter "c", and inserted or deleted hyphens. Only the registrant of a name in a bundle can register other names from the same bundle.

5.3. CL

The .CL TLD (for Chile) publishes a policy page [CLPOLICY]. It has published an IDN table for the Latin script in [IANAIDN]. The policy says that variants are not considered for registration.

5.4. CN

The .CN TLD (for China) publishes its policy as [RFC4713]. It has published an IDN table for the Chinese language in [IANAIDN]. The policy says that variants are "added into the zone file", presumably as NS records.
5.5. ES

The .ES TLD (for Spain) publishes an IDN Area page [ESIDN]. It allows ten accented vowels, U+00E7 ("c with cedilla"), U+00F1 ("n with tilde"), and the Catalan "ela geminada" written as two l’s separated by a U+00B7 ("middle dot"). There are no published IDN tables, and there appears to be no variant policy.

5.6. EU

The .EU TLD (for Europe) publishes a policy page [EUPOLICY]. It has published IDN tables for three scripts in [IANAIDN]. There appears to be no variant policy.

5.7. GR

The .GR TLD (for Greece) publishes a policy page [GRPOLICY] and an FAQ [GRFAQ]. The policy says that all variants of a name under .GR are assigned to the domain owner, with the zone pointing the NS records of all the variants to the name server of the "main form" of the registered name. The FAQ says that domain names in Greek characters are inserted in the zone using their non-punctuated form in Punycode and that the punctuated form is associated with the non-punctuated with a DNAME record. It does not publish IDN tables in [IANAIDN].

5.8. IL

The .IL TLD (for Israel) publishes a policy page [ILPOLICY]. It has published an IDN table for the Hebrew language in [IANAIDN]. There is no variant policy.

5.9. IR

The .IR TLD (for Iran) publishes a policy page [IRPOLICY]. It has published an IDN table for the Persian language in [IANAIDN]. The IDN table says that it will block registration of variants. However, the policy document says that no IDNs can be registered in .IR.

5.10. JP

The .JP TLD (for Japan) publishes a policy page [JPPOLICY]. It has published an IDN table for the Japanese language in [IANAIDN]. Each code point in that table defines no variants, which means there are no variants in registration or resolution.
5.11. KR

The .KR TLD (for South Korea) appears to only publish its policy as an IDN table for the Korean language in [IANAIDN]. The policy in that table does not discuss variants.

5.12. MY

The .MY TLD (for Malaysia) appears to only publish its policy as an IDN table for the Jawi language in [IANAIDN]; however, IANA lists that as a table for "Malay (macrolanguage)". The policy in that table does not discuss variants.

5.13. NZ

The .NZ TLD (for New Zealand) publishes a policy page [NZPOLICY]. It has published IDN tables for the Latin script in [IANAIDN]. The policy does not discuss variants.

5.14. PL

The .PL TLD (for Poland) publishes a policy page [PLPOLICY]. It has published IDN tables for numerous European languages in [IANAIDN]. The policy says that it will block registration of "look-alike" variants.

5.15. RS

The .RS TLD (for Serbia) publishes a policy page [RSPOLICY]. It has published IDN tables for the Serbian language, and the Latin script, in [IANAIDN]. The policy does not discuss variants.

5.16. RU

The .RU TLD (for Russia) appears to only publish its policy as an IDN table for the Russian language in [IANAIDN]. The policy in that table does not discuss variants.

5.17. SA

The .SA TLD (for Saudi Arabia) publishes a policy page [SAPOLICY]. It has published an IDN table for the Arabic language in [IANAIDN]. The policy permits the registration of variants, but it is not clear whether others can register names with variants if the owner of a name has not registered them.
5.18. SE

The .SE TLD (for Sweden) publishes a policy page [SEPOLICY]. It has published IDN tables for the Swedish and Yiddish languages, and the Latin script, in [IANAIDN]. The policy does not discuss variants.

5.19. TW

The .TW TLD (for Taiwan) appears to only publish its policy as an IDN table for the Chinese language in [IANAIDN]. The policy in that table does not discuss variants.

5.20. UA

The .UA TLD (for Ukraine) publishes a policy page [UAPOLICY]. It has published an IDN table for the Cyrillic script in [IANAIDN]. The policy does not discuss variants.

5.21. VE

The .VE TLD (for Venezuela) appears to only publish its policy as an IDN table for the Spanish language in [IANAIDN]. The policy in that table does not discuss variants.

5.22. XN--90A3AC

The .XN--90A3AC TLD (for Serbia) (U+0441 U+0440 U+0431) publishes a policy page [RSIDNPOLICY]. It has published IDN tables for the Cyrillic script in [IANAIDN]. The policy does not discuss variants.

5.23. XN--MGBERP4A5D4AR

The .XN--MGBERP4A5D4AR TLD (for Saudi Arabia) (U+0627 U+0644 U+0633 U+0639 U+0648 U+062F U+064A U+0629) appears to only publish its policy as an IDN table for the Arabic script in [IANAIDN]. The policy permits the registration of variants, but it is not clear whether others can register names with variants if the owner of a name has not registered them.

6. Acknowledgements

Many people contributed to this document, particularly Nacho Amadoz, Marc Blanchet, Michelle Coon, Jordi Iparraguirre, Frederico A. C. Neves, Vaggelis Segredakis, Doron Shikmoni, Andrew Sullivan, Dennis Tan, and Joseph Yee.
7. Security Considerations

There are many potential security considerations for various methods of dealing with IDN variants. However, this document is only a catalog of current variant policies and does not address whether they are good or bad ideas from a security standpoint. The documents cited in the Terminology section have a little discussion of security considerations for IDN variants.

8. Informative References


[VRSNCHAR] Verisign, "Character Variants",
domain-name-services/domain-information-center/
idn-resources/character-variants/index.xhtml>.

[VRSNLANG] Verisign, "Scripts and Languages",
domain-name-services/domain-information-center/
idn-resources/scripts-languages/index.xhtml>.

[VRSNRULES] Verisign, "Registration Rules",
domain-name-services/domain-information-center/
idn-code-points/registration-rules/index.xhtml>.

Authors’ Addresses

John Levine
Taughannock Networks
EMail: standards@taugh.com

Paul Hoffman
VPN Consortium
EMail: paul.hoffman@vpnc.org