IDN Variant Issues Project
Updated Definitions Document

15 October 2011

Purpose of this Document

This updated Definitions document has been developed by the ICANN IDN Variant Issues Project (VIP) team to provide a starting point for the set of definitions of the terms to be used in the *final issues report.* This document is prepared for discussion, and it is expected that its content will change on the basis of discussion prior to, during, and after the Dakar ICANN meeting in October, 2011. In particular, the team expects that discussion will reveal where further synthesis is possible or desirable (or both), and that the terms in need of definition will be determined in part by the shape the final report takes.

This document uses definitions from many documents that have been developed outside ICANN. The primary documents used are:

- Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework (RFC 5890[1])
- Terminology Used in Internationalization in the IETF (RFC 6365[2])
- The Unicode standard including the standard annexes[3]
- Each of the six case study team reports (available at: http://www.icann.org/en/announcements/announcement-4-03oct11-en.htm).

Many of the terms in the script-specific sections are copied directly from their respective reports, and contain cross-references and discussion that makes sense only in the context of those reports. They are included here for convenience, and copied verbatim in order to avoid the accidental introduction of any errors in meaning.

Methodology

In this section, we describe our proposed methodology for synthesizing these definitions.

1) First, we separate terms into two categories: general terms and case study specific terms. General terms refer to terms that are generic and applicable to all cases (e.g. U-label, A-label). Case study specific terms refer to those terms that are used only in a specific case study.

[1] See http://tools.ietf.org/html/rfc5890.

[2] See http://tools.ietf.org/html/rfc6365

[3] See http://unicode.org/versions/Unicode6.0.0/

2) For general terms, if it is defined in RFC 5890, RFC 6365, the Unicode Standard, we copy the relevant definitions from those RFCs or simply refer to them.

3) For case study specific terms, we copy directly from the definition sections of the case study report, if any.

Discussions and input needed by the coordination team for the following scenarios:

4) There are situations where a term defined and used by one case study but can be applicable to other case studies. This is an issue that needs to be explored further.

5) There are situations where a term defined by separate teams means essentially the same thing. We recommend the coordination team harmonize it. There are also situations where one term defined by separate teams has different meanings; this is more problematic and needs careful consideration.

6) There are also situations where a term defined in one case study may be confusing for other studies. This issue also needs to be carefully thought through.

Format of the Definitions in this Document

In the body of this document, the source for the definition is shown in angle brackets, such as “<Unicode>”, “<RFC6365>”. Many definitions are shown as "<IDNVIP>", which means that the definitions were crafted originally for this document. For case study specific teams, they are shown as <Arabic|Chinese|Cyrillic|Devanagari|Greek|Latin-VIP>. Editorial notes, particularly about problematic definitions, are in square brackets ([]).

For some terms, there are commentary and examples after the definitions. In those cases, the part before the angle brackets is the definition that comes from the original source, and the part after the angle brackets is commentary that is not a definition (such as an example or further exposition).

Generic Terms

Abstract Character: A unit of information used for the organization, control, or representation of textual data. <Unicode Standard, section 3.4, D7>

A-label: An ASCII-Compatible Encoding form of an IDNA-valid string. It must be a complete label: IDNA is defined for labels, not for parts of them and not for complete domain names. This means, by definition, that every A-label will begin with the IDNA ACE prefix, "xn--", followed by a string that is a valid output of the Punycode algorithm (RFC 3492) and hence a maximum of 59 ASCII characters in length. The prefix and string together must conform to all requirements for a label that can be stored in the DNS including conformance to the rules for LDH labels (See RFC 5890, Section RFC 2.3.1). If and only if a string meeting the above requirements can be decoded into a U-label is it an A-label. <RFC 5890>

Allocation: In a DNS context, the first step on the way to Delegation. A registry (the parent side) is managing a zone. The registry makes an administrative association between a string and some entity that requests the string, making the string a (candidate) label inside the zone, and a candidate for delegation. Allocation does not affect the DNS itself at all. <IDNVIP>

Assigned Code Point: A mapping from an Abstract Character to a particular Code Point in the code space. <Unicode Standard, section 2.4>

Code Point: A value in the Unicode code space. The meaning here is restricted to meaning D10 in the Unicode Standard, section 3.4.

Delegation: In a DNS context, the act of entering parent-side NS (nameserver) records in a zone, thereby creating a subordinate namespace with its own SOA (start of authority) record. See RFC 1034 for detailed discussion of how the DNS name space is broken up into zones. <IDNVIP>

Domain: [This definition is missing.]

Font[4]: A collection of glyphs used for the visual depiction of character data. A font is often associated with a set of parameters (for example, size, posture, weight, and serifness), which, when set to particular values, generate a collection of imagable glyphs. <UNICODE>

Fundamental Label: A U-label that consists only of Valid Code Points. In practice, this is the U-label requested to be registered. <IDNVIP> [This term has some serious problems with it, and likely needs to be redefined. In particular, as written it also applies to all the variants.]

Fundamental TLD: The Fundamental Label form of a Variant TLD Set. <IDNVIP>

[4] Source: http://unicode.org/glossary/#F

Glyph[5]: (1) An abstract form that represents one or more glyph images. (2) A synonym for glyph image. In displaying Unicode character data, one or more glyphs may be selected to depict a particular character. These glyphs are selected by a rendering engine during composition and layout processing. <UNICODE>

Glyph Image[6]: The actual concrete image of a glyph representation having been rasterized or otherwise imaged onto some display surface. <UNICODE>

IDNA Symmetry Constraint: A-label/U-label transformation must be symmetric: an A-label A1 must be capable of being produced by conversion from a U-label U1, and that U-label U1 must be capable of being produced by conversion from A-label A1. <RFC 5890>

Language Character Repertoire: A set of Code Points identified by some identifier (such as a tag for identifying language as defined in RFC 5646). The definition of the Language Character Repertoire is ideally performed in a way appropriate to some community of language users, and might colloquially be understood as "the characters used to write a language". In most cases, all the Code Points in a Language Character Repertoire will come from the same Script Table. <IDNVIP>

Script Table: A Script Table is a table of Unicode Code Points all having the same script property value. <Unicode Standard Annex #24>

U-label: An IDNA-valid string of Unicode Code Points, in Normalization Form C (NFC) and including at least one non-ASCII character, expressed in a standard Unicode Encoding Form (such as UTF-8). It is also subject to the constraints about permitted characters that are specified in Section 4.2 of RFC 5891 and the rules in the Sections 2 and 3 of RFC 5892, the Bidi constraints in RFC 5893 if it contains any character from scripts that are written right to left, and the IDNA Symmetry Constraint. <RFC 5890>

Writing style[7]: Conventions of writing the same script in different styles. Different communities using the script may find text in different writing styles difficult to read and possibly unintelligible. For example, the Perso-Arabic Nastaliq writing style and the Arabic Naskh writing style both use the Arabic script but have very different renderings and are not mutually comprehensible. Writing styles may

[5] Source: http://unicode.org/glossary/#G
[6] Source: http://unicode.org/glossary/#G
[7] Source: http://www.rfc-editor.org/rfc/rfc6365.txt

have significant impact on internationalization; for example, the Nastaliq writing style requires significantly more line height than Naskh writing style. <RFC6365>

Case Specific Terms

Arabic

Arabic Letter[8]: Characters that are used to write Arabic script based languages and used to write words.

Non-Joining Characters[9]: Those characters that do not connect to letters before or after them; e.g. U+0621 LETTER HAMZA, U+0674 HIGH HAMZA, and U+200C ZWNJ.

Right-Joining Characters[10]: Those characters that connect to the letter before them; e.g. all letters based on ALEF, REH, DAL, and WAW, and a few other letters.

Dual-Joining Characters[11]: Those characters that connect to the letters before and after them; e.g. all other Arabic letters than those listed above.

Join-Causing Characters[12]: Those characters that connect to the letters before and after them, but do not change shape themselves; i.e. only U+200D ZWJ and U+0640 TATWEEL. With respect to those categories, Arabic Script Letters could be defined as follows:

Non-Joining Letters: The group of non-joining characters which are letters (by Unicode's definition); i.e. U+0621 LETTER HAMZA and U+0674 HIGH HAMZA.

Joining Letter: The union of Right-Joining Letters and Dual-Joining Letters which cursively join with letters following them.

[8] Motivated from: http://unicode.org/glossary/#L
[9] Unicode book, Chapter8, table 8-3, page 248 of latest edition
[10] Unicode book, Chapter8, table 8-3, page 248 of latest edition
[11] Unicode book, Chapter8, table 8-3, page 248 of latest edition
[12] Unicode book, Chapter8, table 8-3, page 248 of latest edition

Right-Joining Letters: The group of right-joining characters which are letters; i.e. all letters based on ALEF, REH, DAL, and WAW, and a few other letters.

Dual-Joining Letters: The group of dual-joining characters which are letters; i.e. all other Arabic letters.

Arabic-Indic Digits[13]: Forms of decimal digits commonly used along with Arabic script and comprised of two sets of digits. The set <U+0660-9> is commonly used in Arabic-speaking world, while the set <U+06F0-9>, often referred to as Eastern Arabic-Indic, is used in Iran and Pakistan. Although European digits (1, 2, 3,…) derive historically from these forms, they are visually distinct and are coded separately. (Arabic-Indic digits are sometimes called Indic numerals; however, this nomenclature leads to confusion with the digits currently used with the scripts of India.)

Arabic Digits[14]: The term "Arabic digits" may mean either the digits in the Arabic script (see above and Arabic-Indic digits) or the ordinary ASCII digits. When the term "Arabic digits" is used in Unicode specifications, it means Arabic-Indic digits.

Combining Marks[15]: A commonly used synonym for combining character; a character with the General Category of Combining Mark (M).

Label Valid Character: An Abstract Character which can be used to form a label, and can be a Letter, Digit or another type.

Ligature[16]: A single glyph representing a combination of one or more Arabic Letters. This means that the isolated form of a character can be considered a ligature. It's worth noting that this is different from the Unicode use of the term.

Form of a Letter - Arabic Script: A Letter in Arabic Script can occur in up to four different forms within a ligature. These include the following:

Isolated form:
It is the standalone form of a Letter, i.e. when the letter does not join with any other letter, forming a single letter.

[13] Motivated from: http://unicode.org/glossary/#arabic_digits

[14] Motivated from: http://unicode.org/glossary/#arabic_digits

[15] Source: http://unicode.org/glossary/#C

[16] Motivated from: http://unicode.org/glossary/#L

Initial form:
It is the form of a right-joining-letter when it occurs in the beginning of a ligature, joined with at least one more letter after it, to form a ligature.

Medial form:
It is the form of a right-joining-letter when it occurs in the middle of a ligature, joined with at least one letter on either side, to form a ligature.

Final form:
It is the form of a joining-letter when it occurs at the end of a ligature, joined with at least one more character before it, to form a ligature.

Valid Label: A Label (U-Label or A-Label) which is valid as per the Label Generation Policy.

Valid TLD Label: A Label which is valid for Top Level Domain as per a TLD Label Generation Policy.

Protocol Valid Code Point[17]**:** A Code Point that is allowed to be used in IDNs. Code points with this property value are permitted for general use in IDNs. However, that a label consists only of code points that have this property value does not imply that the label can be used in any given zone.

Label Valid Code Point[18]**:** The subset of Protocol Valid Code Point listed in the Label Generation Policy, and which may be used to form a label.

TLD Label Valid Code Point[19]**:** The subset of Label Valid Code Point that may be used to form a TLD label.

Arabic Script Character Variant[20]**:** A Label Valid Character which is replaceable with another Label Valid Character within a label, as defined by a Label Generation Policy. The relationship is symmetric in Arabic script.

[17] Motivated from http://www.rfc-editor.org/rfc/rfc5892.txt

[18] Motivated from Draft Definitions for the ICANN Variant Issues Project Document 'Valid Code Point'

[19] Motivated from Draft Definitions for the ICANN Variant Issues Project Document 'Valid Code Point'

[20] Motivated from Draft Definitions for the ICANN Variant Issues Project Document 'Character Variant'

Label Generation Policy[21]: A formal specification that can be used to formulate or validate a label and determine whether two labels can be considered distinct for allocation. If two labels are not considered distinct for allocation, as per the policy, they are referred to as variants of each other. Variants are symmetric in Arabic script.

TLD Label Generation Policy: A formal specification that can be used to formulate or validate a TLD label and determine whether two TLD labels can be considered distinct for allocation.

Arabic Script Label Generation Policy: Policy specified to generate labels for Arabic script. For Arabic script, this would include at least a list of Protocol Valid Code Points allowed in forming labels, their Character Variants, additional Label formation constraints/rules and meta information (e.g. including script, owner, version, date, etc.).

Arabic Script TLD Label Generation Policy: Policy specified to generate TLD labels for Arabic script.

Variant Character Set[22]: The set of code points consisting of a Valid Code Point and all of its variants.

Script Table[23]**:** A Script Table is a table of Unicode Code Points all having the same script property value. See Unicode Standard Annex #24.

Variant Label: A U-label considered a variant of a Fundamental Label as per the Label Generation Policy.

[21] This term is being suggested as a replacement for "Language Table" as the latter has been found inappropriate as it "Language Table" is not limited to a single language and may represent many languages or an entire script, and "table" represents an implementation level decision (as eventually an XML based or other specification could also be used. Also, language table seems to be a binding, if not insufficient, mechanism to specify additional rules/constraints needed and does not clearly specify method or contents for meta information required to promote unambiguous use and reuse of the information.

[22] Motivated from Draft Definitions for the ICANN Variant Issues Project Document 'Variant Character Collection'

[23] Motivated from Draft Definitions for the ICANN Variant Issues Project Document 'Script Table'

Fundamental Label[24]: A Valid Label which is, in practice, the label received by the registry to be allocated and from which a registry or registrar may generate Variant Labels.

Activated Variant Label: A Variant Label that is activated by a registry.

Allocated Variant Label: A Variant Label that is allocated by a registry.

Reserved Variant Label: A Variant Label that is reserved by a registry without allocation but may be allocated on request

Blocked Variant Label: A Variant Label that is blocked by a registry (to avoid a conflict) and is not allowed to be allocated

Variant Label Set[25]: A set of U-labels consisting of one Fundamental Label and its zero or more Variant Labels.

Activated Variant Label Subset: The subset of Variant Label Set that is activated, or alternatively, the set containing the Fundamental Label and all its Activated Variants.

Allocated Variant Label Subset: The subset of Variant Label Set that is allocated, or alternatively, the set containing the Fundamental Label and all its Allocated Variants.

Reserved Variant Label Subset: The subset of Variant Label Set that is reserved, or alternatively, the set containing all the Reserved Variants of a Fundamental Label (and the Fundamental Label, if it is not activated).

Blocked Variant Label Subset: The subset of Variant Label Set that is blocked, or alternatively, the set containing all the Blocked Variants of a Fundamental Label.

Activation: The process of making a domain name resolvable.

Reservation: In Arabic Script IDN variants context, this is the process of having an unallocated variant label which relates to a Fundamental label that is allocated.

[24] Motivated from Draft Definitions for the ICANN Variant Issues Project Document 'Fundamental Label'

[25] Motivated from Draft Definitions for the ICANN Variant Issues Project Document 'Variant Label Set'

Blocking: In Arabic Script IDN variants context, this is the process of having a variant label not allowed for allocation to anyone as long as its Fundamental label is allocated.

Chinese

Han Script Variant: Characters with different visual forms but with the same pronunciations and with the same meanings as the corresponding official forms in the given language contexts. For more details please refer to the section 5 of this report.

Official Form: In different country/region, the government specifies "official forms" for a set of general use Hanzi. In Mainland China, they are called normalized Hanzi (规范字 U+89C4 U+8303 U+5B57), and in Taiwan, they are called orthographic Hanzi (正體字 U+6B63 U+9AD4 U+5B57).

CJK Characters: "CJK characters" are characters commonly used in the Chinese, Japanese, or Korean languages, including but not limited to those defined in the Unicode Standard as ASCII (U+0020 to U+007F), Han ideographs (U+3400 to U+9FAF and U+20000 to U+2A6DF and U+2A700 to U+2B73F and U+2B740 to U+2B81F), Bopomofo (U+3100 to U+312F and U+31A0 to U+31BF), Kana (U+3040 to U+30FF), Jamo (U+1100 to 11FF and U+3130 to U+318F), Hangul (U+AC00 to U+D7AF and U+3130 to U+318F), Kangxi Radicals(U+2F00 to U+2FDF), CJK Radicals Supplement (U+2E80 to U+2EFF), and the respective compatibility forms.

CJK Unified Ideograph: An ideograph is a graphic symbol that represents an idea. Chinese Hanzi, Japanese Kanji and Korean Hanja are often referred to as ideographs. Since 1990, tens of thousands of Chinese Hanzi, Japanese Kanji and Korean Hanja have been merged into CJK Unified Ideographs and their Extension in ISO/IEC 10646 and Unicode. In this document, if not otherwise specified, the term "ideograph" means a CJK Unified Ideograph.

Language Variant Table: The key mechanism used by current domain registries for calculating Chinese variant labels is a three-column table, called a Language Variant Table, designated for each language permitted to be registered in the zone. Those columns are known, respectively, as "Valid Code Point," "Preferred Variant," and "Character Variant," and are defined separately below. In this document, “LVT” and “Variant Table” are used as short forms for “Language Variant Table.” IANA maintains the list of variant tables for the Chinese Script. Appendix C provides some examples.

Valid Code Point: In a Language Variant Table, a “Valid Code Point” is an entry on the list of code points that is permitted to be registered for that language. Any other code points, or any string containing them, will be rejected by this specification. The Valid Code Point list appears as the first column of the Language Variant Table.

Preferred Variant: In a Language Variant Table, a “Preferred Variant” is an entry on the list of code points corresponding to each Valid Code Point and providing possible substitutions for it. These substitutions are "preferred" in the sense that the variant labels generated using them are normally registered in the zone file, or “activated.” The Preferred Code Points appear in column two of the Language Variant Table. “Preferred Code Point” is used interchangeably with this term.

Character Variant: In a Language Variant Table, a “Character Variant” is an entry on the second list of code points corresponding to each Valid Code Point and providing possible substitutions for it. Unlike the Preferred Variants, substitutions based on Character Variants are normally reserved but not actually registered (or “activated”). Character Variants appear in column 3 of the Language Variant Table. The term “Code Point Variant” is used interchangeably with this term.

Zone Variant: A “Zone Variant” is either a Preferred or Character Variant Label that is actually to be entered (registered) into the DNS, that is, into the zone file for the relevant zone. Zone Variants are also referred to as Zone Variant Labels, active labels, or Activated Labels.

Internationalised Domain Label (IDL): The term “Internationalized Domain Label” or “IDL” will be used instead of the more general term “IDN” or its equivalents. This is the string of characters of the domain name being applied for and has been validated as suitable for inclusion in the DNS zone file. In the case of an IDN TLD, the IDL is simply the string of characters of the TLD being applied for and has passed the evaluation.

IDL Package: An “IDL Package” is a collection of IDLs as determined by the guidelines in RFC 3743. All labels in the package are "reserved", meaning they cannot be registered by anyone other than the holder of the Package. These reserved IDLs may be "activated", meaning they are actually entered into a zone file as a "Zone Variant". The IDL Package also contains the language tag. The IDL and its variant labels form a single, atomic unit, however, not all labels in the package are active.

Language Tag: Language tags, as defined in RFC 5646, are used to help identify languages, whether spoken, written, signed, or otherwise signaled, for the purpose of communication. This includes constructed and artificial languages but excludes languages not intended primarily for human communication, such as programming languages.

Cyrillic

Alternate Names: Two names are alternates of one another just in case, for a namespace starting with one, the namespace starting with the other is isomorphic to the first, subject to the usual DNS loose consistency strictures. In the current DNS, there are 2 different techniques for this. The first is aliasing: CNAME, DNAME, and other such techniques redirect a name or a tree, effectively substituting one label for another during DNS lookup. The second is by using provisioning constraints, such that an underlying provisioning system always effects a change in all of the alternate names whenever that change is effected in one of the alternates. A fuller discussion of this topic is included for information in Appendix B.

Composite-character variants: Abstract Characters that do not have a single assigned code point assigned, but can be represented by multiple code points.

Domain Name Blocking Policy: Refers to a policy that has effect of certain domain names in a TLD registry becoming unavailable for allocation (for example, due to implementation of variant-related policies).

Domain Name Bundling: Registration technique that makes multiple domain names share all registration parameters (such as creation/expiration date, associated name servers etc.) except the domain name itself. Changes to any of these registration parameters should normally take effect on all the domain names in a bundle.

Reserved name: A name set aside for a potential allocation to a particular registrant (or TLD registry in the case of TLDs in the root). The name is not allocated, but could be if/when certain conditions are met.

Greek

Homograph: A word that shares the same written form with a word of the same or different script but may have different meanings. The same could apply for any string of letters, even if not a word. The characters that make this possible between scripts are mentioned as homograph characters in the document (e.g. the Latin A and the Greek A).

Homophone: Two different words that are pronounced the same. The same could apply for any string of letters, even if not a word.

Greeklish (Source: Wikipedia, http://en.wikipedia.org/wiki/Greeklish, text in Italics): a portmanteau of the words Greek and English, also known as Grenglish, Latinoellinika/Λατινοελληνικά or ASCII Greek, is the Greek language written using the Latin alphabet. Unlike standardized systems of Romanization of Greek, as used internationally for purposes such as rendering Greek proper names or place names, or for bibliographic purposes, the term Greeklish mainly refers to informal, ad-hoc practices of writing Greek text in environments where the use of the Greek alphabet is technically impossible or cumbersome, especially in electronic media. Greeklish was commonly used on the Internet when Greek people communicate by forum, e-mail, IRC, instant messaging and occasionally on SMS, mainly because older operating systems didn't have the ability to write in Greek, or in a Unicode form like UTF-8. Nowadays most Greek-related content appears in native Greek.

Aliased name: A domain name that has been aliased with one or more names under the concept of Name Aliasing. The technical solution that is currently available for aliasing a domain name to another is the use of a CNAME or a DNAME record in a zone file, essentially mapping a domain name or a DNS tree to another. Voices for other technical solutions have been raised over the last few years but without any results.

Name Aliasing: The abstract concept of two or more domain names "behaving as one" by Policy or technical means. This concept has still unresolved issues, definitional, technical and political. Currently this concept is technically served by the use of CNAME and DNAME records in a zone file, allowing for the aliasing of a domain name or a DNS tree to another. With the introduction of the concept of variants, arguments have been raised for the necessity of the presentation of a mechanism that would allow the users to keep in sync domain name trees but without the limitations of DNAME and CNAME. Experts have discussed the use of a CNAME+DNAME kind of record but without concluding on the advantages and disadvantages such a record would present. Issues with the use of DNSSEC, and the inability to address what “behaving as one domain name” stands for in the real world have stalled the discussions on this issue.

Bundling Domain Names: The registry policy of registering certain domain names as a set, depending on select characteristics (e.g. homograph domain names). In some cases, bundling is used for name aliasing purposes.

Tonos: Greek accent mark, acute accent (Greek Tonos, U+0384)

Dialytika (diaeresis): Greek accent mark (appears on the letters “ϊ” (e.g. Greek small letter iota with dialytika, U+03CA) and “ϋ” (e.g. Greek small letter upsilon with dialytika, U+03CB) to show that a pair of vowel letters is pronounced separately, rather than as a diphthong – see http://en.wikipedia.org/wiki/Diphthong). It can also be combined with tonos over the same letters, Greek small letter iota with dialytika and tonos, U+0390 and Greek small letter upsilon with dialytika and tonos, U+03B0.

Katharevousa: (Greek: Καθαρεύουσα, [kaθaˈ revusa], lit. "puristic [language]"), is a form of the Greek language conceived in the early 19th century as a compromise between Ancient Greek and the Modern Greek of the time, with a vocabulary largely based on ancient forms, but a much-simplified grammar. Originally, it was widely used both for literary and official purposes, though seldom in daily language. In the 20th century, it was increasingly used for official and formal purposes, until Dimotiki became the official language of Greece in 1976 (Source: Wikipedia, http://en.wikipedia.org/wiki/Katharevousa).

Dimotiki: (Greek: δημοτική [γλώσσα] [ðimotiˈ ci], "[language] of the people") is the modern vernacular form of the Greek language. The term has been in use since 1818. Demotiki refers particularly to the form of the language that evolved naturally from ancient Greek, in opposition to the artificially archaic Katharevousa, which was the official standard until 1976. The two complemented each other in a typical example of diglossia until the resolution of the Greek language question in favour of Demotiki (Source: Wikipedia, http://en.wikipedia.org/wiki/Dimotiki).