# Draft Questions for Variant Case Study Teams

June 13, 2011

## 1. Background

ICANN is working to identify how to best provide for delegation of multiple Internationalized Domain Name (IDN) Generic Top-Level Domains (gTLDs) that are Variant TLDs. To support this work, an issues report is being developed to identify the issues related to "what needs to be done with the evaluation, possible delegation, allocation and operation of gTLDs containing variant characters IDNs … in order to facilitate the development of workable approaches to the deployment of gTLDs containing variant characters IDNs."[1]

This work is being undertaken by performing a number of language- and script-specific case studies. By having a number of representative groups focus on variant issues specific to their language, the issues can first be explored unencumbered by the possible requirements of other language groups. Once these case study groups develop consensus on their specific requirements, it is intended that as far as possible the issues will then be synthesized together and an overarching issues report developed.

**This document lists some framing questions for case study teams. It is intended be used as a starting point to stimulate discussions. It is *not* meant to be a complete list of questions.**

## 2. Scope of the Work and Issues Reports

The following is the scope of work:

1. Identify appropriate terminology for the various concepts and requirements, ensuring such terms are accurate and vetted with appropriate technical and linguistic communities and are used consistently throughout the project to improve the dialogue among participants;
2. Identify the requirements considering (a) linguistic accuracy, (2) technical feasibility, (c) usability, (d) accessibility, and (e) security and stability.

The following items are *not* within the scope of the work:

3. Determine the circumstances (where they exist) where certain types of IDN variant TLDs might be eligible for delegation;

[1] See Board resolution: (http://www.icann.org/en/minutes/resolutions-25sep10-en.htm#2.5

4. Analyze and arrive at rules where possible, or guidelines where rules are not possible, that address the challenges of working with IDN variant TLDs outlined in task 2;
5. Arrive at rules and guidelines, both in the registry operational requirement area and the technical implementation area;
6. Determine the responsibilities of TLD operators who would be responsible for managing such delegated IDN variant TLDs;
7. Determine what kind of compliance programs may be necessary to ensure that IDN variant TLDs operate according to the arrived at rules and guidelines;
8. Identify viable and sustainable outreach mechanisms to communicate and interact with the community on the issues report

.

Tasks (3) through (8) will be the focus of follow-on projects for ICANN policy development, implementation guidelines produced by ICANN staff in consultation with the community, and relevant technical work by other interested organizations.

The individual case study teams should not concern themselves with developing solutions that are designed to accommodate the needs of other language groups. The key focus of the case study groups is to come up with an agreeable definition of that language group's needs, in terms of linguistic requirements, user experience, and so on. It will be the role of the final issues group to harmonize the differing requirements, terminology and other aspects of the case study groups.

The case study approach proceeds from the assumption that IDNA (sometimes called "IDNA2008"; see RFC 5890, RFC 5891, RFC 5892, RFC 5893, and RFC 5894) is the basic technology to be used for IDNs. That assumption entails the use of Unicode. Individual case study teams may raise objections to these assumptions. If your team has such objections, please outline the degree to which you believe waiting for an alternative to IDNA is acceptable.

## 3. Questions

These are a starting set of questions for the case study team to consider.

### 3.1 Definitions

As many of the terms used in variant discussions to date have different definitions to different people (including the meaning of the terms "variant" itself), each group is asked to settle on common terminology, and use that terminology consistently across its discussions. Some initial common terminology is defined in a separate document to these questions (refer to "Draft Definitions for IDN Variant Issues Project"). If the terms there defined are not appropriate, we prefer another (different) term should be defined (instead of altering the definition of the defined term). The goal of this restriction is to minimize the difficulty later during the harmonization step. It is acceptable to use subscripts,

neologisms, or any other mechanism to call out the difference between the new term and the term as defined in the initial set.

- Are the initial definitions of Variant Character and Variant Character Label sufficient for your case, or is another definition required (provide examples as well as references for your definition that are vetted by linguistic and technical communities)?

- Are the other definitions sufficient for your case, or are other definitions required (provide examples as well as references for your definitions that are vetted by linguistic and technical communities)?

- Is it possible to specify a Language Character Repertoire for the language and script community (or communities) under consideration?

- Are there other terms needed? If so, what are they, and how are they to be defined?

## 3.2 Basic Character Questions

- Are there Variant Characters, in the Language Character Repertoire?

- Are the variant characters the same across all languages written in the script?

- In a set of variant characters, are the relationships between all the Variant Members symmetric? For example, in German it is possible to use the characters "ss" as a variant for the character ß (U+00DF, LATIN SMALL LETTER SHARP S). But not everything that contains "ss" may be spelled with "ß". This is an example of a case where the Variant Members do not have a symmetric relationship.

## 3.3 General questions about domain labels

These questions relate to the use of a U-label at any position in an IDN.

- Are variant characters required for U-labels in the script or can they be avoided?

- Apart from the protocol limits, is there any limit on the number of characters in a U-label that are members of a Variant Character Collection?

- Is it acceptable that every case of a Character Variant Label or Preferred Variant Label be replaced by the corresponding Fundamental Label?

- Is there a requirement that it be possible for more than one member of a Variant Label Set to be delegated, and if so, why?

- What are the repercussions if a member of the Variant Label Set is required to be delegated, but cannot be?

- How should the rules for handling variant characters be expressed in the relevant Language Variant Table?

- Can a single Language Variant Table be standardized for the entire script (covering all languages expressed in the script)? If not, what effect does that have?

- Is information about Character Variant Labels required to be known by parties other than the registry in which the U-labels are registered (and, where appropriate, IANA)? If so, how is such information to be communicated?

- Are all members of the Variant Label Set to be allocated to the one authority? If not, what are the implications?

- What is the end user expectation when using a Fundamental Label, Preferred Variant Label, or Character Variant TLD? Consider more than just URLs in web browsers. What is the effect of using a variant in an email address? In an ENUM context? When cutting and pasting Internet names into non-network applications like word processors, and then back out again? On business cards?

- Does the end user need to know about Variant Label Sets and if so how is this information to be made available to the end user?

- Are there end user confusion issues when some member of the Variant Label Set that the end user might reasonably expect to "work" does not (e.g. a Reserved Variant TLD)? How serious is it? Will users adapt easily and naturally to such cases, or will this be a persistent problem for a long time?

- What is the system administrator expectation when configuring for Variant Label Sets (need to know all variants / simplicity / etc.)? Consider the effects on a system administrator operating the DNS, and operating servers that need to know the names by which they are known (e.g. mail servers, web servers, and so on).
- What are the DNSSEC requirements if more than one U-label in the Variant Label Set needs to be delegated? What effects might this have on delegations lower in the tree?

- What are the SSL/TLS requirements if more than one label in the Variant Label Set needs to be delegated? Consider all cases: X.509 certificates, SSH keys in the DNS, TLS keys in the DNS (as proposed by the DANE IETF Working Group), and so on. What effects will this have on subordinate domains in the DNS?

- What are the general Internet protocol requirements (i.e. beyond common protocols such as HTTP) for the use of members of the Variant Label Set?

### 3.3.1 Variant Label Sets at the TLD

Consider the above questions with respect to the Variant TLD Set. It is not enough to consider just the effects of the variant approach on labels immediately beneath them (i.e. at the second level): the policies at a TLD will affect every domain beneath them too. What are the effects of different variant behavior in ccTLDs (which are perhaps understood to be directed at a national and well-known linguistic population) and in gTLDs (where the population maybe presumed to be global)?

### 3.3.2 Variant Label Sets at lower levels

Consider the above questions with respect to U-labels lower in the DNS tree. Might the effects be different? Scope differences and differences in kind are both important. Is it reasonable to permit multi-script domains at lower points in the tree when it is not reasonable to permit as much at the top level? Why or why not? What are the practical effects of differing policies between a parent domain and one of its descendants?

## 3.4 Scripts with Identical Characters

- Are domain name labels in scripts with identical characters to those in the TLD script permitted, and, if not, how are these to be blocked and which script is given priority and why?

## 3.5 Other Questions

- What needs to be done with the evaluation of IDN variants?
- What needs to be done with the allocation of IDN variants?
- What needs to be done with delegation of IDN variants?
- What needs to be done with operation of IDN variants?
- What Policy should ICANN set on language tables? Should ICANN only permit one language table for a given script, or it may be desirable to have multiple language tables for a given script? If so, on what basis we are to make these exceptions? What should happen if multiple tables submitted for a given language/script does not agree with each other?
- What operational criteria ICANN should specify for applicants of variant TLDs? Should we look at their technical capabilities, if so, what should we look at? Do we look at their registration policies, if so, what should we look at? What's the scope of ICANN's review of variant strings, and of the operator's capability?

- Following up on the previous question, what contract terms should ICANN impose on operators? And how should ICANN enforce those contracts?

- What is ICANN's process for evaluating these variants? Several questions in this category:

    - What should the evaluation process look like? Specifically how should the guidebook be amended?
    - What's the appropriate fee model. Is it the same for for people who already put variants on the list? What is the appropriate and fair structure for application fee; should it change depending on number of variants applied?
    - Each registry pays quarterly fees to ICANN. With variant TLDs, if an operator has multiple variants, should they pay more? What is the fair approach here?

- What should be the difference in consideration of IDN ccTLDs and IDN gTLDs, if any?

**Appendix I:**

# Draft Definitions for the The ICANN Variant Issues Project

## Background

The ICANN Variant Issues Project is investigating issues around variants and the global DNS, with particular attention to the issues for top level Internationalized Domain Names using IDNA2008 (and, therefore, Unicode).

The Project provides the following definitions as a starting definition both as part of the definition of terms used in the Questions for Case Study Teams, and as terms to be used across all Case Study Teams.

The Project asks that Case Study Teams use the terms only as defined below. If a term is defined in a way that the Team finds undesirable, or if there is not a term for something the Team needs to discuss, the Project prefers the Team to define a new term (if only by adding a subscript to the term as defined here), so that it is always clear what a given term means. The goal of this restriction is to minimize the difficulty later during the harmonization step.

To avoid reinventing the wheels, we borrowed / reused some definitions from the following documents:

- Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework (RFC 5890)
- Terminology Used in Internationalization in the IETF (draft-ietf-appsawg-rfc3536bis-01)
- The Unicode standard including the standard annexes
- Joint Engineering Team (JET) Guidelines for Internationalized Domain Names (IDN) Registration and Administration for Chinese, Japanese, and Korean (RFC 3743)

## Definitions

**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 5390, 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 label inside the zone, and a candidate for delegation. Allocation does not affect the DNS itself at all.

**Assigned Code Point:** A mapping from an Abstract Character to a particular Code Point in the code space. See Unicode Standard, section 2.4. Not to be confused with Valid Code Point.

**Character Variant**: In a Language Variant Table, a 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 Variants" is used interchangeably with this term. (RFC 3743)

**Character Variant Label**: A U-label generated by use of Character Variants. This definition differs from that in RFC 3743 by specifying “U-label” rather than “label”.

**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.

**Fundamental Label**: A U-label that consists only of Valid Code Points. In practice, this is the U-label requested to be registered.

**Fundamental TLD**: The Fundamental Label form of a Variant TLD Set.

**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.

**Language Variant Table**: A three-column table for each Language Character Repertoire permitted to be registered in a zone. The columns are known, respectively, as "Valid Code Point", "Preferred Variant", and "Character Variant", which are defined separately. (This definition differs from RFC 3743 in the subsitution of Language Character Repertoire for “language”.) Note that in the rest of this document "Table" and "Variant Table" are *not* used as short forms for Language Variant Table, as they are in RFC 3743. Note also that it is logically possible a U-label would be consistent with more than one Language Variant Table. What to do in such a case is a matter of registry policy.

**Preferred Variant**: In a Language Variant Table, a 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 2 of the Language Variant Table. "Preferred Code Point" is used interchangeably with this term. (RFC 3743)

**Preferred Variant Label**: A U-label generated by use of Preferred Variants. This definition differs from that in RFC 3743 by specifying “U-label” rather than “label”.

**Preferred Variant TLD**: The Preferred Variant Label form(s) of a Variant TLD Set.

**Reserved Variant TLD**: The Character Variant Label form(s) of a Variant TLD Set.

**Script Table:** A Script Table is a table of Unicode Code Points all having the same script property value. See 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)

**Valid Code Point**: In a Language Variant Table, the list of Code Points that is permitted for that language. Any other Code Points, or any string containing them, will be

rejected. The Valid Code Point list appears as the first column of the Language Variant Table. (RFC 3743) Note that Valid Code Points are always both Assigned Code Points and Variant Members.

**Variant Character Collection**: All the characters listed in a single row of a Language Variant Table, as any of Valid Code Point, Preferred Variant, or Character Variant. (RFC 3743) It is important to recognize that the relationship may not be reciprocal (that is, if *foo* is a Valid Code Point and *bar* is a Character Variant, that does not mean that *foo* is a Character Variant for Valid Code Point *bar*).

**Variant Label Set**: A set of U-labels consisting of one Fundamental Label, zero or more Preferrred Variant Labels, and zero or more Character Variant Labels.

**Variant Members**: Code Points that appear in a Language Variant Table. The code point may appear in any of the Valid Code Point, Preferred Variant, or Character Variant positions.

**Variant TLD**: A Variant Domain Name Label corresponding to an A-label that appears or is intended to appear immediately below the root in the global DNS. Note that this definition includes TLDs that do not actually exist in the DNS at a given point in time. More informally, a Variant Domain Name Label that appears or intended to appear immediately below the DNS root. Because the actual labels in the DNS are all A-labels, this informal use is not strictly true; but because A-labels and U-labels are symmetric, it amounts to the same thing.

**Variant TLD Set**: A set of Variant TLDs consisting of one Fundamental Label, zero or more Preferred Variant Labels, and zero or more Character Variant Labels.

### Other items to note

The definitions of variants above is different from other meanings of "variant" sometimes heard in an ICANN context. Those alternate meanings for the term "variant" include the following:

- Different abstract characters (see Unicode Standard section 3.4 D7) that may be visually confusable. Example: Arabic-Indic Digit Seven (U+0667, "٧") and Latin Small Letter V (U+0076, "v").
- Characters that are normally graphically identical, but that have different Assigned Code Points. Example: Cyrillic Small Letter A (U+0430, "а") and Latin Small Letter A (U+0061, "a").
- Orthographic differences within a language. Many languages have alternate choices of spellings or spellings that differ by locale. Users of those languages generally recognize the spellings as equivalent. An example is "color" and "colour" in English.

Of course, different choices of fonts or other forms of visual representations, including handwriting, can cause any pair of characters in any language to look more or less alike. For example, while Greek Lower Case Alpha (U+03B1) is usually considered visually distinguishable from the Latin and Cyrillic Lower Case A characters mentioned above, there are typefaces for basic Latin characters that make the lower case A look almost like Greek Lower Case Alpha. None of these other meanings of "variant" are the kind of variant discussed in the ICANN Variant Issues Project.