Report on Chinese Variants in Internationalized Top-Level Domains

James Seng

on the behalf of the Chinese Case Study Team

24th Oct 2011

Chinese Case Study Team

Name	Role
Xiaodong Lee	Case Study Coordinator
Chris Dillon	Team Member
Hong Xue	Team Member
James Seng	Team Member
Jian Zhang	Team Member
Jonathan Shea	Team Member
Joseph Yee	Team Member
June Seo	Team Member
Shian-Shyong Tseng	Team Member
Wei Wang	Team Member
Yangwoo Ko	Team Member
Yoshiro Yoneya	Team Member
Zhoucai Zhang	Team Member
Edmon Chung	Observer
Yang Yu	Observer
Steve Sheng	Case Study Liaison
Francisco Arias	Subject Matter Expert (Registry Operations)
Kim Davies	Subject Matter Expert (Security)
Nicholas Ostler	Subject Matter Expert (Linguistics)
Andrew Sullivan	Subject Matter Export (Protocols)

Scope of Work

- Focus on Chinese Variants at the TLD
  - But we acknowledge the effect on the lower levels;
  - consider user expectations and consistency between the top-level and lower-level

- Focus on Chinese Variant
  - But we recognized Chinese variant *might* have implications for Japanese and Korean users who also use Han script;
  - Includes experts from Japan and Korea within the team who identifies potential impact for the Japanese and Korean community

- Focus on Unicode
  - Anything that is not included in Unicode is considered out of scope

Chinese/Han Script

- Chinese uses Han script (Hanzi/漢字/汉字) which consist of tens of thousands of characters, of which several thousands are in common use.
- Japanese also uses Han script (Kanji/漢字) in addition to Hiragana and Katakana script and Korean also uses Han script (Hanja/漢字/한자) in historical context but are currently using Hanguel script (한글).
- Chinese hanzi, Japanese Kanji and Korean Hanjia are often referred to as ideograph, a graphic symbol that represents an idea.
  - All Han script has a unique concept/meaning and could be combined with others to extend or form new concept/meaning.
- In Unicode, Han script is unified into “CJK Unified Ideographs” (about 26,000 characters) and Extension B, C and D (about 44,000 characters).

Chinese Hanzi

- Chinese Hanzi originate from pictographs, which later evolves into ideographs over several thousands of years.
  - e.g, the ideograph for the concept of “hill” is 山 still bears some resembles three peak of a hill.

- In 1964 and 1986, the National Language Committee introduces more than 2,000 Simplified Hanzi as official form in Mainland China.
  - 350 Simplified Component: ‘treasure’ 寶 -> 宝
  - 132 Complex Component: ‘dragon’ 龍 -> 龙
  - 14 Radials: ‘concept of talking’ 言 -> 讠
  - 1,753 Simplified Hanzi: ‘deaf’ 聾 -> 聋 ‘lesson’ 課 -> 课

- As a result, the Chinese language has two writing system: Simplified Chinese (SC) and Traditional Chinese (TC), both using the same Han script albeit a different subset.
  - Simplified Chinese is the official form in Mainland China and Singapore and Traditional Chinese is the more commonly use in Taiwan, Hong Kong, Macau. Singapore, Malaysia and other oversea Chinese community uses a mixture of SC and TC.

- Most importantly, SC and TC have the same meaning, the same pronunciation, and are typical variants.
  - But they are not always 1-to-1: 發 and 髮 have the same simplified form 发.

Japanese Kanji

- Japanese Kanji (漢字) were imported from China and used as ideographic characters. There are over 6,000 Kanji used in Japan, 2,000 of which are commonly used (Jōyō Kanji/常用漢字).

- Some of them are in a simplified form or “new character form”(新字体) derived from the “old character form” (旧字体).
  - The Kanji simplification is done independently from Chinese and therefore the mappings are different from Chinese Hanzi.

- New and old character form are recognized as variants in general concept except when uses to express nouns such as name of persons and places, the new and old form are considered distinct.

- As domain names are often uses to express names, it is appropriate to distinguish old and new forms as different independent characters instead of handling them as variants.

Korean Hanja

- Korean Hanja (漢字/한자) were also import from China and were widely used as ideographic character in historical context.

- Hanguel (한글) is a systemic phonetic script designed by King Sejong in the 15th century and replaced Hanja as the preferred writing system.
  - Though many Korean vocabularies are derived from Hanja, they are now written in Hanguel.

- Today, Hanja is no longer widely used in South Korea and a law was enacted on 14th April 2011 that all government documents can only be written in Hanguel, unless allowed by Presidential decree.

- Registry operator for .KR (Korea Internet & Security Agency), does not allow Hanja in their IDN policy and have no intention of allowing the use of Hanja.

Define Chinese Variants

"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"

- In Chinese, two type of variants
  - Regional standard writing system: Simplified and Traditional Chinese
  - Generic variants: Several Chinese characters have slightly different visual form and are considered the same but given different code point in Unicode. e.g. 户/戶" (U+6237 / U+6236) and "黃/黃" (U+9EC4 / U+9EC3)

- What they are not
  - Different spellings/form, presentation in multiple script, translations of a string, transliterations of ring are not considered variants.

Chinese User Expectations

- Because Chinese variants have the same pronunciation and the same meaning as its official form, Chinese users regard them as interchangeable.

- If Chinese variants are not permitted to be delegated, the coherent recognition of Chinese characters will be seriously jeopardized.
  - CNNIC has 320,000 Chinese IDNs and 77% have variants form, TWNIC has 40,000 Chinese IDNs and 83% have variants form and HKNIC has 24,000 Chinese IDNs and 85.5% have variants form.

- If only one variant is delegated, Chinese users would be deprive of using the variant most comfortable to them.
  - CNNIC experiences with .中国/.中國 shows over 10% of the DNS queries are for TC form

- If variants are operated by two independent registries, this could cause bad user expectations and potentially security issues as it may lead to phishing attack.

Chinese & Japanese Variants Handling at the TLD

- For Japanese, there is no need for the consideration for variants in the context of names as new and old form are considered distinct.

- 学会.jp and 學會.jp would be considered two different domain names in Japan but 学会.cn and 學會.cn would be considered variants domain names in China.
  - The TLD provides certain contextual indictor and would be able to co-exists

- At the TLD, .学会 and .學會 would be viewed as TC/SC variants TLD by Chinese users and must be delegated while the rest of the variants to be reserved.

- For Japanese, taking a conservative approach, an application for .学会 would results the following variants sets {学會 ,學會, 學会, 孝會, 孝会} to be reserved.

Language Variant Tables

- The need for Chinese Language Variant Tables for the Root Zone
- Whether IDN variants at TLD level should be based on language or script
- Considerations for a process to define the root Variant Tables
- The standard for Variant Tables

Language Variant Tables

The need for Chinese Language Variant Tables for the Root Zone

a) ICANN allows applicants to submit a Variant Table together with their application

The problem with this approach is that ICANN will have to verify the accuracy of the Variant Table from each applicant.

b) ICANN allows applicants to submit the variants of the TLD with their application

The problem with this approach is that with no Variant Table, there is no technical mechanism to for ICANN to verify whether the requested variant is reasonable or accurate.

c) ICANN adopts a single unified variant table for the root zone

The problem with this approach is the difficulty for ICANN to adopt a single Variant Table.

Language Variant Tables

Whether IDN variants at TLD level should be based on language or script

- If IDN variants at the TLD level were generated based on language, then there would be different variants (or no variants) of the same IDN TLD string depending on the language. This would result in the most accurate variant handling for Chinese script with least false positives.

- If IDN variants at the TLD level are generated based on CJK script wide tables, then there would be same variants for the same IDN TLD string, regardless of the language. However, this would result in wider false positives, whereby there may be variants been reserved even though such variants may not be linguistically accurate and correct in the applied language.

- Thus, we prefer variants to be generated based on language but, since this is apparently impossible at the root, we can accept a script based system.

Language Variant Tables

Considerations for a process to define the root Variant Tables

a) May have more than one Variant Tables for the root

b) Reference to Standards defined by an International recognized Standard Body, such as Unicode Consortium or ISO/IEC JTC1/SC2.

c) Relying on long established practices within the ICANN community, such as Chinese Domain Name Consortium (CDNC)

d) Calling upon industry and linguistic experts

e) IANA to maintain the Variant Tables

Language Variant Tables

The standard for Variant Tables

- The earliest work on variant tables for the Chinese script is RFC 3743 for Chinese, Japanese and Korean domain name registration guideline.

- The Chinese community, building on the concepts and principles of RFC 3743, defined RFC 4713, which is a more specific recommendation for Chinese domain name registration and administration.

- RFC 4290 also derives its basic principles and concepts from RFC 3743 and attempts to make a generic recommendation for internationalized domain name registration. It retains many of the concepts without going into the specifics of the algorithm for generation of variants.

Evaluation, Allocation, Delegation and Operation of Chinese IDN Variant TLDs

- Evaluation Issues (when ICANN evaluate the TLD application), such as string similarity, conflict with geographic names, discovery of variants, etc

- Contention/Objection/Dispute Issues (when ICANN rules that there is a contention for the applied-for gTLD, or when objections or disputes are received for the new TLD)

- Allocation Issues (when ICANN decides which string(s) is/are to be allocated for the new TLD)

- Delegation Issues (when ICANN decides whether the applicant is the right authority to be delegated the new TLD string(s) and whether there is a fee involved)

Evaluation, Allocation, Delegation and Operation of Chinese IDN Variant TLDs

Evaluation Issues (when ICANN evaluate the TLD application)

- Evaluation Panels - it is desirable that experts with knowledge and experience of linguistics, especially the Chinese language, be included in the Panel to look at string similarity, identify conflicts with geographic names, and verify the list of discovered variant labels for the applied for IDL.

- Discovery of Chinese Variant Labels - It is advisable that ICANN will proactively identify the IDL packages of all applied-for gTLDs in Chinese because the applicant may not have provided the variant labels of the applied-for gTLD in an automated process to ensure consistency and evaluation process.

- Technical and Financial Readiness - The applicant should be aware of and ready to address issues arising from having to administer additional TLDs which are Preferred Variant Labels of the applied for TLD.

Evaluation, Allocation, Delegation and Operation of Chinese IDN Variant TLDs

Contention/Objection/Dispute Issues (when ICANN rules that there is a contention for the applied-for gTLD, or when objections or disputes are received for the new TLD)

The presence of IDN variants will lead to more scenarios to be considered when deciding whether there is a string contention.

- What if the Application A is in contention with the variant of Application B?
- What if variant of Application A is in contention with variant of Application B?
- What if A and B of the above two cases happen to be in different languages?
- What if the variant is a minor case that none of applicants care about?

IDN variants should also be taken into account in Objection Procedures, particularly in "string confusion objection" and "legal right objection."

Evaluation, Allocation, Delegation and Operation of Chinese IDN Variant TLDs

Allocation Issues (when ICANN decides which string(s) is/are to be allocated for the new TLD)

It is advisable that the whole IDL Package that passes the Evaluation be allocated. The fact that all labels in the IDL Package are allocated does not mean that all labels in the IDL Package have to be put into the zone file (i.e. activated)

Evaluation, Allocation, Delegation and Operation of Chinese IDN Variant TLDs

Delegation Issues (when ICANN decides whether the applicant is the right authority to be delegated the new TLD string(s) and whether there is a fee involved)

The following 3 sets of TLDs should be delegated together:

1. Applied-for IDL
2. Preferred Variant in Simplified Chinese (based on zh-CN table)
3. Preferred Variant in Traditional Chinese (based on zh-TW table)

Issues

- Delegation of Variant Labels
- Security and Stability of the DNS
- Registry Fees to ICANN
- Contractual Provisioning Requirements
- Delegating a reserved variant TLD at a later date
- Un-delegating a Variant Label which was delegated previously

Evaluation, Allocation, Delegation and Operation of Chinese IDN Variant TLDs

Operation Issues (operational consideration after the new TLDs have been delegated)

- Impact to Registry Operations: DNS Resolution, DNSSEC, Shared Registration System, WHOIS, Data Escrow, Trademark Clearing House, etc

- Impact to Dispute Resolutions
  - Should new gTLD trademark measures, specifically Post-Delegate Dispute Resolution Policy (PDDRP), take into account the IDN variants of both the disputed domain names and pertinent right holders' ' trademarks?
  - Should the trademark clearinghouse operator accept the submission of variant forms of a trademark in IDN characters?
  - When someone raises a dispute of a domain name under the new TLD, should the same
  - domain name and/or its preferred variant labels under the variant TLDs (if any) be also considered together or separately? What if these TLDs are administered by different registries? What if the variant labels of that domain name are held by different registrants?
  - Should new gTLD trademark measures, specifically Uniform Rapid Suspension Policy (URS), take into account the IDN variants of both the disputed domain names and pertinent right holder' 's trademarks?

Considerations

- IANA : Impact on Variant Tables, Variant Resolution Mechanism, IDN WHOIS standards
- Root Server operators: Impact on the number of entries in the root zone
- Registries: Evaluation, Allocation, Delegation and Operation
- Registrars: Fees/Charges for Variant? EPP changes?
- Registrants: registration policy and management of variants
- Domain Name related providers: how would the upper layer applications work with variants (SMTP, HTTP, IMAP etc)? Adjustment to infrastructure
- Software/Application Vendors: configuration for variants handling? Tools for variant handling?

Conclusion

- The Chinese case study team considers that language variant tables are a critical element in selecting eligible top-level domains, and therefore the policy on how those tables are developed is very important.

- The variant issues will impact the TLD application, evaluation, allocation, delegation and operation. IANA, root server operators, registries, registrars, registrants, domain name related service providers, software/ application providers and other stakeholders need to consider the variant issues when they adopt the variant TLDs.