Potential Issues in Chinese IDN Variant TLDs

Prof. Xiaodong LEE

China Internet Network Information Center(CNNIC)
Chinese Academy of Sciences (CAS)

On behalf of Chinese Domain Name Consortium (CDNC)

Contents

- What's Chinese Variant?
- Example: Chinese IDN Variant
- What the requirement of Chinese IDN Variant TLDs?
- CJK Considerations
- Potential Issues

CNNIC
中国互联网络信息中心
China Internet Network Information Center

What's Chinese Variant

- Today, Chinese-language users throughout the world are more closely connected to China and Chinese culture than ever before.
  - According to a Report issued by Chinese Tourism Bureau in 2009, a total of 104.45 million passengers went into Mainland China from Taiwan, Hong Kong, and Macau; in the meantime, the number of tourists from mainland China to those three regions reached 29 million.
  - Furthermore, apart from 1.4 billion Chinese users in Mainland China, Taiwan, Hong Kong and Macau, there are over 48 million Chinese users living in other parts of the World.
- Chinese has two written forms
  - Simplified Chinese (SC), which is used primarily in Mainland China and Singapore
  - Traditional Chinese (TC), which is used primarily in Taiwan, Hong Kong, other Southeast Asian countries, and communities of Chinese origin in other countries (most of whose population emigrated a generation or more ago).

What's Chinese Variant

- In a world where SC and TC are recognized as interchangeable, Chinese language users expect to be able to access Chinese information seamlessly and with optimal readability and usability.
  - Millions of Chinese users use both SC and TC in their daily communications. Any attempt to separate SC from TC could, at the very least, create user confusion and, at worst, result in the marginalization of millions of users.
  - Moreover, allowing both SC and TC—and regarding them as identical—offers tremendous convenience to Chinese users whose keyboards, input methods, and sometimes display methods support only one or the other.
- San Francisco
  - It was said that over 200 thousand Chinese people living in Bay Area, most of them are speaking Cantonese and writing TC, but in recent 30 years, more and more are speaking Mandarin and writing SC.

CNNIC
中国互联网络信息中心
China Internet Network Information Center

What's Chinese Variant

- Simplified characters correspond to more complex Traditional characters . These corresponding sets of characters, which are referred to in RFC 3743 and successor documents as character variants, share the same meaning and pronunciation, but they do not look the same.
- In Chinese, many characters have a corresponding variant, specifically in the CDNC Chinese Character Table
  - Please refer to RFC 3743 and, more specifically, to Chinese rules of RFC 4713.
  - 19520 Chinese characters are opened for Chinese domain name registration, and 7890 Chinese characters have one or more variants
    - http://www.iana.org/domains/idn-tables/tables/cn_zh-cn_4.0.html
    - http://www.iana.org/domains/idn-tables/tables/tw_zh-tw_4.0.1.html
    - In Unicode 5.1, there are 74,394 CJK Unified Ideograph
  - NOTE: Chinese Variant issue doesn't mean only SC and TC equivalence

中国互联网络信息中心
China Internet Network Information Center

Example: Chinese IDN Variant

- In Chinese, the strings “中国银行.中国” and “中國銀行.中國” are variant to each other, which means “Bank of China.China”
  - If registered to two different registrants, a user in Mainland China entering a domain name in SC could be directed to one site,
  - while another user in Taiwan entering what they perceive to be the same domain name in TC is directed to a different site.
  - This may create confusion for end users and would most likely create bad user experiences; it could possibly even invite serious phishing attacks.
- Hence, ensuring that both the SC and TC versions of a domain name are registered to the same registrant will avoid confusion to the end user.
- Chinese Variant string has SC-only, TC-only and Mixed-SC-TC ones. Mixed-SC-TC ones mostly never used by users, except for typo.

CNNIC
中国互联网络信息中心
China Internet Network Information Center

Example: Chinese IDN Variant

- If, for example, only “中国银行.中国” (SC-only BankofChina.China) is allowed to be registered, a traditional script user will not be able to input the domain name and will, hence, be excluded from accessing the website.
  - Permitting only one of the names may place some users of the same language, belonging to the same culture, at a significant disadvantage; worse still, it could lead to segregation of populations that are part of that language and cultural group.
- As a further example
  - While Hong Kong and Macau uses TC predominantly, growing usage of SC is experienced.
  - Even in Mainland China, so many companies and organizations use the Traditional Chinese as their brand name and LOGO.

中国科学院 CHINESE ACADEMY OF SCIENCES

中国银行 BANK OF CHINA

CNNIC 中国互联网络信息中心 China Internet Network Information Center

What's the requirement of Chinese IDN Variant TLDs

- If an applicant applies a Chinese IDN TLD, mostly it will be a SC-only or TC-only form.
  - The language tag should be provided, it means that the applicant should tell which community he want to serve.
  - If Chinese, both SC-only and TC-only forms MUST be added into the Root Servers, otherwise not.
  - Mixed-SC-TC forms and other variants SHOULD be reserved for this applicant, and be blocked by any future other application to avoid confusing
    - Except for some special strings which should be considered separately
  - See RFC3743 (JET) and RFC4713 (CDNC) for further information
- What's the issue for CJK, which are using HAN character

CNNIC
中国互联网络信息中心
China Internet Network Information Center

CJK Considerations

- Almost all modern scholarship on writing system categorizes scripts that are actively used today (other than by scholars) and that have a very long history (a thousand years or more) into two major categories. Those categories are based both on origin and on how the writing system works and consist of:
  - the Han-derived “Chinese” system
  - and the collection of scripts known as “alphabetic” or “phonetic”.
- Chinese script contains some phonetic elements, but is considered to be primarily ideographic
  - or logographic, or symbolic, depending on authors and whether distinctions are being made among characters within the script
- Chinese is the only ideographic/symbolic script in use by native-speaker populations today.
  - For the Chinese script, Japan has actually been using simplified character forms for many years, much longer than those forms have been official in China, but does not alternate them with traditional forms.
  - Insofar as Chinese characters (known as Hanja) are likely to be used in Korea at all (The official characters are Hangul), Korea uses traditional forms only.

CNNIC
中国互联网络信息中心
China Internet Network Information Center

CJK Considerations

- Since Japanese and Korean populations do not need or use variants, CDNs and their variants can only be used by the Chinese-language community.
  - The registries associated with CDNC worked with the Japanese and Korean registries to develop the JET Guidelines described in RFC 3743.
  - Japan and Korea use a compatible model for examining the relevant characters, but do not use variants and so are not affected by the main elements of this proposal.
  - Please see the ".KR" Korea Character Table and the ".JP" Japanese Character Table to understand the requirements of them, which are link separately to
    - http://www.iana.org/domains/idn-tables/tables/kr_ko-kr_1.0.html
    - http://www.iana.org/domains/idn-tables/tables/jp_ja-jp_1.2.html.
- So, while the script is the same and evaluations should be carried out on the same principles (for example, the Simplified Chinese form of a Korean registration might be treated as potentially confusing and blocked), only the Chinese language needs SC/TC variants as paired TLD delegations for the Chinese script.

中国互联网络信息中心
China Internet Network Information Center

Potential Issues

- String Similarity Evaluation
  - 3000 necessary, 6000 popular, 19520 open, other 50000 potential
    - What's the limitation?
    - How many is enough?
  - Different forms, but similar meanings
    - 吃，食（eat）
  - Different meanings, but similar forms
    - 日，曰（Sun，Say）
- Delegation Combination (Better than nothing)
  - Paired Delegation, with improvement on administration
  - IETF Long term solution, waiting for ten years?

中国互联网络信息中心
China Internet Network Information Center

One world, One Internet!

THANKS!

LEE@CNNIC.CN

CNNIC
中国互联网络信息中心
China Internet Network Information Center