Network Working Group
Internet-Draft
Expires: Nov 21, 2002

XiaoDong LEE
Kenny Huang
Erin Chen
Xiang DENG
YanFeng WANG

Chinese Name String in Search-based access model for the DNS

draft-xdlee-cnnamestr-01.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Content

1. Abstract
2. Terminology
3. CNS equivalence
4. Requirements
5. Solution suggested
6. Encoding
7. Security Considerations
8. Authors' Addresses
9. Acknowledgements
10. References

1. Abstract

There are many requirements of developing internationalized and human-readable Internet identifiers/names now, thereby there are many systems based on DNS technology to meet such requirements. John C. Klensin has proposed a three-layer search-based access model for the DNS [DNSSEARCH]; this paper is only to explain some related problems mentioned in John C. Klensin's proposal. Especially it focuses on Traditional and Simplified Chinese problems and some other special

Chinese requirements.

The ultimate goal for any kinds of search-based access system is to help users to access network resources in more natural ways, which have different meaning for different user groups. On the premise of respecting Chinese user's language convention, it is very important for a valuable and human-friendly system to deal with traditional and simplified Chinese equivalence problems.

2. Terminology

The key words "SHALL", "REQUIRED", "SHOULD", "RECOMMENDED", "MUST", and "MAY" in this paper are to be interpreted as described in [RFC2119].

In order to describe the problem simply, we define these terminologies first.

"TC" is an abbreviation for Traditional Chinese.

"SC" is an abbreviation for Simplified Chinese.

"CNS" is defined as an acronym of Chinese Name String that is the most important facet, name string mentioned in [DNSSEARCH], which contains at least one Chinese character. As to the scope of Chinese character, please refer to ISO/IEC 10646-1:2000(E) [second edition 2000-09-15], if one character is marked "C and G-Hanzi-T", it MUST be a Chinese character, such definition does not mean it is not the character of other countries that use HAN ideograph.

"TC-only CNS" is a CNS that all characters of it are TC characters.

"SC-only CNS" is a CNS that all characters of it are SC characters.

"Mixed-use TC and SC CNS" is a CNS of which at least one traditional and one simplified Chinese character appear in all characters.

3. CNS equivalence

The TC/SC equivalence problem is very complex and difficult to solve perfectly, please refer to [CTCC], nevertheless, there are mainly three categories of single TC/SC character equivalence, so we should solve these problems respectively and one by one, after solving these three kinds of problems, most of the TC/SC problems will be solved, and the result will be acceptable for most Chinese users.

a) One to one

E.g. U+98A8 (TC, "the wind") can be mapped to U+98CE (SC, the wind)
U+5099 (TC, to prepare) can be mapped to U+5907 (SC, to prepare)
U+908A (TC, a side) can be mapped to U+8FB9 (SC, a side)

b) One to many

E.g. U+6FF1 (TC, the shore) can be mapped to U+6EE8,U+6D5C (SC, the shore)
U+53C3 (TC, three, to take part in) can be mapped to U+53C2 (SC, to take part in) U+53C1 (SC, three)
U+58DF (TC, a ridge or walkway in a field) can be mapped to U+5784,U+5785

(SC, a ridge or walkway in a field)
c) Many to one
E.g. U+85F9,U+8B6A (TC, friendly) can be mapped to U+853C (SC, friendly)
U+5225 (TC, to leave), U+5F46 (TC, to awkward) can be mapped to U+522B (SC, to leave, to awkward)
U+93DF (TC, a shovel), U+5277 (TC, a shovel) can be mapped to U+94F2 (SC, a shovel)
But as to the equivalent problem of CNS, it is a combination of above three categories, so it is more complex than single character, but we could process it one character by one character.

4. Requirements

These requirements SHOULD be considered for any system supported Chinese name string.
a) TC and SC CNS equivalent matching
SC is derived from TC, and Chinese people use both SC and TC. So Chinese people think that TC CNS is equivalent with its corresponding SC forms, so any implementation should meet such requirement.
b) Mixed TC and SC CNS cause an exponential problem
If we want to ensure a CNS in both TC/SC forms to be resolved correctly, we could register all its forms, but along with the length of label, an exponential problem will occur. Most of Chinese character variants are daily used. An ordinary Chinese Name String may have dozens of, hundreds of, even thousands of TC/SC variants. That is unreasonable for users to register, and uneasy for administrators to manage, and complex for system to resolve. No matter which kind of search-based access system, flat or hierarchy, or central-controlled, and so on, it is not reasonable for any administration to process these thousands of name strings un-automatically.
c) Some other special requirement
As you know, there are many conventional differences between Chinese and English. Such as of name string sequence. English people could write "Minneapolis, Minnesota" to represent a location, but Chinese people would like to write as "Minnesota, Minneapolis". So if we permit search-based access system to use sequence attributes to represent delegation or hierarchy, such kind of special requirement should be satisfied.

5. Solution suggested

As mentioned in [DNSSEARCH], there are many challenges in doing traditional and simplified Chinese equivalence, because HAN character is not only used in China, but also in other countries, mostly in Asia. To be emphasized firstly, no method could solve traditional and simplified Chinese equivalence perfectly and correctly, and up to now, the best algorithm is only able to achieve about 99%, rather than 100%. So maybe that is the reason why no consensus has been made in IDN WG.

Because we have two facets in search layer two, language and country code/ geographical location, which will be very useful to solve most of the problems. Based on these two facets, system with certain language and country code could pick appropriate rules to do traditional and

simplified Chinese equivalence without any impact on other languages and countries.

In Mainland China, as to "One to One" and "Many to One", we could convert all these TC character into SC character, and then save SC-only CNS into database for Chinese name string resolving. But as to "One to Many", it maybe based on context, the system may handle this in artificial intelligent method, it is a pity that even the best artificial intelligent algorithm cannot solve this conversion completely. As in my opinion, this kind of artificial conversion shouldn't be completed in layer two, which should have affirmative result with some simple facets; these artificial process should be completed in layer three or get user's feedback to make sure which name string he want. User's feedback may be added when doing conversion, or using result cached by last conversion.

E.g.

a) One to one

{[CN] + [zh-cn] + TC} --> {[CN] + [zh-cn] + SC}

b) Many to one

{[CN] + [zh-cn] + TC1/TC2/.../TCn} --> {[CN] + [zh-cn] + SC}

c) One to many

User feedback

{[CN] + [zh-cn] + TC} ------------------> {[CN] + [zh-cn] + SC1/.../SCn}

Finally, all Mixed-use TC and SC CNS should be converted into SC-only CNS before resolving, and only SC-only CNS are stored in resolving database in server. What's more, if we do want to implement "One to Many" conversion in layer two, we could bind the TC CNS with one of its corresponding SC forms with "first come, first use" based on reasonable principle, that is, the binding process should avoid binding two irrelevant CNS and cause meaningless equivalent resolving.

As shown above, Mainland of China could select conversion rules from TC to SC, for TC area, contrary rules from SC to TC could be used. As to this suggestion, user feedback is very important for One to Many conversion, we just provide a solution to resolve CNS correctly, it permit user to input unconventional Mixed-use TC and SC CNS in certain language and country or area, but actually it doesn't happened very often.

Some people suggest to use fuzziness level to determine matching precision, they want user to select which kind of conversion they want, it is not useful to solve TC/SC equivalence problem, I think, traditional and simplified Chinese equivalence problem is not a fuzziness problem as other fuzzy matching problems in search-based access system. Providing fuzziness level Chinese matching will mislead end users, and will cause questionable namespace in layer two. Chinese name string should have same process rules in system level, which should not be based on user intention completely.

6. Encoding

In layer two and layer three or above, as to the encoding of Chinese character, we suggest using UNICODE directly, any additional encoding will increase the system complexity, and it is unreasonable for a long term solution. Of course, local encoding isn't limited, but it should be converted into Unicode encoding before interchanging in internet.

7. Security Considerations

This paper is just a complement document for [DNSSEARCH], so it has same security considerations. TC/SC CNS equivalence problem will not bring any additional security problems into this search-based access model.

8. Authors' Addresses

XiaoDong LEE
Chinese Academy of Sciences, CNNIC
4 South 4th Street, ZhongGuanCun, Beijing 100080
Phone: +86 10 62619750 ext. 3020
E-mail: lee@cnnic.net.cn

Kenny Huang
Taiwan Network Information Center (TWNIC)
4F-2, No.9 Sec. 2, Roosevelt Rd., Taipei, 100 Taiwan
E-mail: huangk@alum.sinica.edu

Erin Chen (also as Yu Hsuan Chen)
Taiwan Network Information Center (TWNIC)
4F-2, No.9 Sec. 2, Roosevelt Rd., Taipei, 100 Taiwan
Phone:: +886 2 23411313 ext. 502
E-mail: erin@twnic.net.tw

Xiang DENG
China Internet Network Information Center(CNNIC)
4 South 4th Street, ZhongGuanCun, Beijing 100080
Phone: +86 10 62619750 ext. 3018
E-mail: deng@cnnic.net.cn

YanFeng WANG
China Internet Network Information Center(CNNIC)
4 South 4th Street, ZhongGuanCun, Beijing 100080
Phone: +86 10 62619750 ext. 3022
E-mail: wyf@cnnic.net.cn

9. Acknowledgements

Thanks for these person's suggestions and efforts.
HuaLin QIAN hlqian@cnnic.net.cn ; CAS, CNNIC
Li-Ming Tseng <tsenglm@csie.ncu.edu.tw>; NCU, TWNIC
Wei MAO mao@cnnic.net.cn ; CNNIC
Wen-Sung Chen <wschen@twnic.net.tw>; TWNIC

10. References

[RFC2119] Scott Bradner, Key words for use in RFCs to Indicate Requirement Levels, March 1997, RFC 2119.

[STD13] Paul Mockapetris, Domain names - implementation and specification, November 1987, STD 13 (RFC 1034 and 1035).

[CTCC] The Pitfalls and Complexities of Chinese to Chinese Conversion Jack Halpern, Jouni Kerman

[ISO10646] ISO/IEC 10646-1:2000. International Standard - Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane.

[Unicode3] The Unicode Consortium, "The Unicode Standard -- Version3.0", ISBN 0-201-61633-5.

[DNSSEARCH] John C. Klensin, "A Search-based access model for the DNS", draft-klensin-dns-search-05.txt, May 2001,