IDN Testbed Deployment at SGNIC in Singapore

Implementation of CDNC Language Table & Development of Tamil Language Table

Presented by:

Edmon Chung, Afilias Ltd.
Tan Yeow Hui, SGNIC

SGNIC Requirements

- Official Languages in Singapore
  - English, Malay, Chinese & Tamil
- Malay
  - Can be represented with LDH (Letters-Digits-Hyphens, i.e. ASCII) and will NOT be tagged
- Chinese
  - Simplified (SC) & Traditional Chinese (TC) Required
  - Use of CDNC (Chinese Domain Name Consortium) Table
- Tamil
  - No Readily Available Language Table

SGNIC IDN Testbed Concept

- Fully Standards Compliant
  - IETF RFCs & ICANN Guidelines
- Testbed 2LD registry used: <IDN>.idn.sg
- 1 Year Testbed Period
- Free Registration through Registrars
- Testbed IDNs Deleted Upon Conclusion
- Chinese Domain Names Basic Policy:
  - Primary Label + Preferred SC + Preferred TC
  - All are automatically “Activated”
  - Allow “Activation” of Other Variants
- Tamil Domain Names
  - No Variant Preparation Required

Study of the Chinese Language Table

- Based on “Registration and Administration Guideline for Chinese Domain Names”
  - www.ietf.org/internet-drafts/draft-xdlee-idn-cdnadmin-01.txt
  - Published: Feb. 15, 2004
- Total Number of Codepoints (Chinese Characters): 19,551
  - LDH: 37 Codepoints
  - Preferred SC: 15,298 (max: 2; single-entry: 19,453 -> 99.7%)
  - Preferred TC: 16,345 (max: 7; single-entry: 18,525 -> 94.6%)
  - Other Variants: 7880 (max: 7; single-entry: 6,333)
- Simplified Chinese (SC) vs. Traditional Chinese (TC)
  - SC=TC: 12,375 (SC=TC=Codepoint: 12,045)
  - SC=TC=Codepoint AND with Other Variants: 411 (w/o: 11,634)
- Does NOT Require Variant: 11,634+37 (Does: 7880)

System Challenges for CDN

- One registration leading to many data entries
  - One Primary Label Registration can lead to >10,000 variants (based on TWNIC's practical experience)
- Storing ALL variants would be problematic
  - Not Economically viable
  - Accelerated exponential increase in lookup time
  - Larger storage required
- Revenue supported transactions & related costs
  - Registration of Primary Label (Domain)
    - Included Automatically Activated "Preferred Variants"
  - Activation of Variants ("activation" = published into zone file)

DB & Matching Consideration

- Two Main Approaches
  - Store all Variants in Database
  - Utilize Variant Index
- Store All
  - Straightforward Implementation
  - Exponential Database Size possibility
- Variant-Index
  - “normalize” Character string to “indexes” for matching
  - Additional overhead per transaction
  - Only “Activated” Labels are stored in DB
- Policy Flexibility Implications
  - Store All: Must Store ALL variants based on policy
  - Variant-Index: Policy can be more flexibly changed

Further Study on CDNC Table

- Objectives:
  - Feasibility of Indexing
  - Choice of Index
  - Uncover Issues with Variant-Index Approach
- Defining an Index
  - Exhaustively group all “related” Codepoints / Variants
  - Lowest Codepoint used as “Index” (purely arbitrary)
- Findings:
  - Total Indexes: 15111 (11634 w/o variants)
  - 3477 Indexes vs. 7880 Codepoints (with variants)
  - 1057 Indexes with “Hidden Variants”

“Hidden Variants”

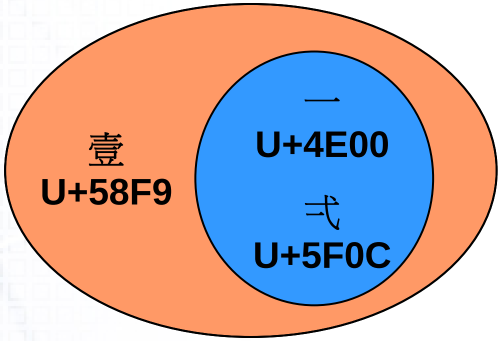
- Caused by Overlaps between Row Entries in the CDNC Table
  - Example: Chinese Character for “One”

Codepoint	Preferred SC	Preferred TC	Other Variants
U+4E00 一	U+4E00 一	U+4E00 一	U+5F0C,U+58F9 弌,壹
U+58F1 壱	U+58F9 壹	U+58F9 壹	U+58F9 壹
U+58F9 壹	U+58F9 壹	U+58F9 壹	U+4E00,U+58F1 一,壱
U+5F0C 弌	U+4E00 一	U+4E00 一	U+4E00 一

Index	Codepoint Set
U+4E00	U+4E00,U+58F1,U+58F9,U+5F0C 一,壱,壹,弌

  - Given: U+58F9 壹
    - TC=SC=Original; Other Variants {U+4E00 一,U+58F1 壱}
    - Hidden Variant: U+5F0C 弌

Variant Set Overlaps

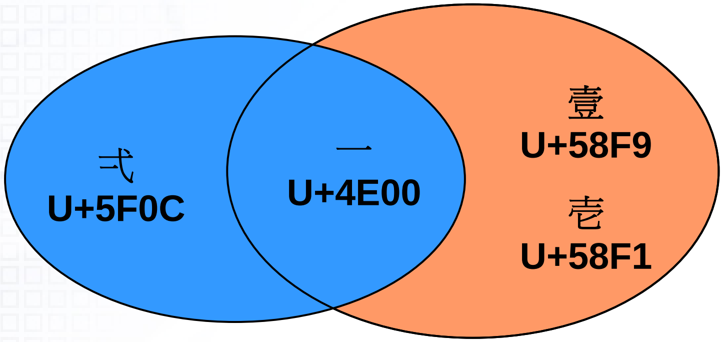


壹
U+58F9

一
U+4E00

弌
U+5F0C

Scenario 1: Variant Subsets



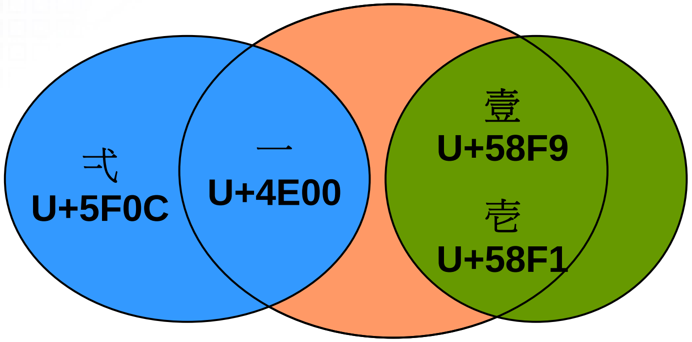
弌
U+5F0C

一
U+4E00

壹
U+58F9

壱
U+58F1

Scenario 2: Overlapping Variant Sets



弌
U+5F0C

一
U+4E00

壹
U+58F9

壱
U+58F1

Scenario 3: Overlap created by 3rd Set

Policy Challenges

- Language Declaration
  - Simplified Chinese / Traditional Chinese / Chinese
- Auto “Activated” Variants
  - Multiple Preferred Characters
  - All Permutations vs. Arbitrary Choice
- WHOIS Display
- “Promotion” of Variants
  - Result of DRP (e.g. Rights to Variant is a Trademark owned by another party)
- “Hidden Variants”
  - Automated Registration Process

CDN Policies for SGNIC Testbed

- Unified / Consolidated SC/TC Table
  - Original; Preferred SC; Preferred TC; Other Variants
  - One Preferred SC and One Preferred TC Automatically “Activated”
  - First codepoint used if multiple Preferred
- Registrant (via Registrar) is allowed to “Activate” and “Deactivate” other variants
  - Variants assume same set of Delegation NS
  - Expiration of Activated Variants are Synched with Primary
  - All registration / activation currently set to $0
- Registrations must not overlap with existing IDNs
  - Adherence to Conservativeness Principle
  - No Promotion of Variants allowed (no DRP within testbed)

Tamil Table Considerations

- No Readily and Publicly Available Tamil Language Table (Unlike CDNC Table)
  - Local (Singapore) Tamil experts representing INFITT, TISC and MINC through SGNIC
  - Indic experts
  - Focused on Tamil Codepage in Unicode (U+0B80 - U+0BFF)
- Identical Codepoints (with numerals)
  - U+0BE7 ௧ (numeral for 1)
  - U+0B95 க (Tamil Character for "KA")
- Allow / Disallow registration of numeral and symbol characters

Tamil Policies for Testbed

- Tamil Language Table:
  - TAMIL SIGN VISARGA = aytham: 0B83
  - Independent vowels: 0B85-0B8A, 0B8E-0B90, 0B92-0B94
  - Consonants: 0B95, 0B99-0B9A, 0B9C, 0B9E-0B9F, 0BA3-0BA4, 0BA8-0BAA, 0BAE-0BB5, 0BB7-0BB9
  - Dependent vowel signs: 0BBE-0BC2, 0BC6-0BC8
  - Two-part dependent vowel signs: 0BCA-0BCC
  - Various signs: 0BCD, 0BD7
- No Numerals or Non-Alphabetic Symbols
- No Variants Preparations Required
- Considering whether to allow LDH or only DH
  - Mixture of English (ASCII) “Letters” with Tamil is not common

Summary

- SGNIC IDN Testbed
  - Standards Compliant (IETF RFCs & ICANN Guidelines)
  - <Chinese/Tamil>.idn.sg (English & Malay => ASCII)
  - Free registrations (All IDNs deleted at the end of testbed)
- Chinese Domain Names
  - CDNC Language Table
  - Registration “Package” {Primary Label + 1 Preferred SC + 1 Preferred TC}
  - Behavior, Expiration and Delegation NS synched with Primary
  - Study on Index Variant and Implications, e.g. “Hidden Variants”
- Tamil Domain Names
  - Tamil Codepage without Numerals & Non Alphabetic Symbols
  - No Variants

Thank You

- Edmon Chung
  - edmon@afilias.info
- Tan Yeow Hui
  - yeowhui@nic.net.sg