Ruling The Root

CJK Rules For The Root Zone

Kenny Huang, Ph.D.  黃勝雄博士

Member, CDNC / CGP
Co-author, RFC3743 IETF
Member, Executive Council, APNIC
Member, Board of Directors, TWNIC
huangksh@gmail.com
2014.Jun
Problem: CJK Is Complicated

Putting CJK labels in the root zone is even more complicated.
Institutionalized Problem Solving: Structure

Unified LGR for the Root Zone

Merge

Integration Panel

C (Chinese) Generation Panel

K (Korean) Generation Panel

J (Japanese) Generation Panel

CJK Coordination Committee
## Constraints for CJK LGR

### Independent Tasks
- Each CJK Panel creates an LGR
- Each LGR includes a repertoire and variants
  - Define labels permission
  - Define variants labels
  - Assign dispositions
    - Allocatable
    - Block

### Coordination Tasks
- If an LGR includes Han characters:
  - The variant *mappings* must agree for all the panels
  - The variant *types* may be different
  - The repertoires may be different

*Presented by Lee Han Chuan & IP, Shanghai 2014 May 29*
Overlap Case Illustration

Chinese LGR

<table>
<thead>
<tr>
<th>Variant</th>
<th>① 壹</th>
<th>② 式</th>
<th>③ 壹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
<td>U58F9</td>
<td>U5F0C</td>
<td>U58F1</td>
</tr>
<tr>
<td>Disposition</td>
<td>allocate</td>
<td>block</td>
<td>block</td>
</tr>
</tbody>
</table>

Japanese LGR

<table>
<thead>
<tr>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
</tr>
<tr>
<td>Disposition</td>
</tr>
</tbody>
</table>

Integrate ?

T

Integrated Root Zone Label Generation Rules

F

Rejected

Generation Panel
# High Level Conflict Strategies

## CJK overlap

<table>
<thead>
<tr>
<th>ID</th>
<th>Strategy</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adopt X Abandon $R_{cjk}$</td>
<td>Permit X</td>
<td>No label rule</td>
</tr>
<tr>
<td>2</td>
<td>Adopt X Intersection $\cap (R_{cjk})$</td>
<td>Permit X Permit $\cap (variants/disp)$</td>
<td>Rules changed</td>
</tr>
<tr>
<td>3</td>
<td>Adopt X Union $\cup (R_{cjk})$</td>
<td>Permit X Permit $\cup (variants/disp)$</td>
<td>Rules changed</td>
</tr>
<tr>
<td>4</td>
<td>Abandon X and $R_{cjk}$</td>
<td>No conflict</td>
<td>Label not available</td>
</tr>
<tr>
<td>5</td>
<td>Adopt rules based on frequency of use</td>
<td>Fair &amp; scientific approach</td>
<td>Rules changed; fairness doesn’t mean appropriate</td>
</tr>
</tbody>
</table>
Unified CJK LGR Illustration

Chinese LGR

<table>
<thead>
<tr>
<th>Variant</th>
<th>壹</th>
<th>弍</th>
<th>壹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
<td>U58F9</td>
<td>U5F0C</td>
<td>U58F1</td>
</tr>
<tr>
<td>Disposition</td>
<td>allocate</td>
<td>block</td>
<td>block</td>
</tr>
</tbody>
</table>

Japanese LGR

<table>
<thead>
<tr>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
</tr>
<tr>
<td>Disposition</td>
</tr>
</tbody>
</table>

Integrated LGR

<table>
<thead>
<tr>
<th>Variant</th>
<th>壹</th>
<th>弍</th>
<th>壹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
<td>U58F9</td>
<td>U5F0C</td>
<td>U58F1</td>
</tr>
<tr>
<td>Disposition</td>
<td>allocate</td>
<td>block</td>
<td>block</td>
</tr>
</tbody>
</table>

Union

Intersection
CJK Integration Methodology

Divide & Conquer (D&C)

Diversified CJK Demands

- C Demands
- J Demands
- K Demands
- Evaluation Method
- LGR
- Constrains
- MSR

Plan and Define

Strategic Direction

Unified CJK Rules

- Root Zone Admin
- CJK Rules
- CJK Repertoire
- Minimal Viable Solution
- Variant Dispositions

Services

- LGR
- Constrains

Diversified CJK Demands

- CJ Overlap
- CK Overlap
- Repertoire
- Repertoire

Split

Merge
Splitting Non-overlapping Code Points From Repertories

CJK Han-overlap in IANA IDN Repository

J-Han : 6356 (JPRS)  K-Han : 0 (KRNIC)

C-Han : 19520 (CNNIC/TWNIC)

Develop Conflict Strategy

No conflict

Problem Domain (Unsolved Overlap) : 6181

unified code points

<table>
<thead>
<tr>
<th>13339</th>
<th>175</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13514</td>
</tr>
</tbody>
</table>

RC  Chinese LGR
Rj  Japanese LGR
RK  Korean LGR
Engineering Design

Computation for Word Usage and Frequency

TC: Apple News
SC: Sina News
JP: Mainichi News

Sample size is statistically significant

C/J overlap code points

Matching

usage

Split unused code points

frequency of use

Split code points of low frequency of use
Splitting Unused Code Points from The Overlap

C / J Overlap Data Set : 6181

total unused : 2739

- \( R_j \) : J only : 203
- \( R_c \) : C only : 1927

C / J usage overlap : 1312

Problem Domain (Unsolved Overlap) : 1312

Unified Code Points:

<table>
<thead>
<tr>
<th></th>
<th>2739</th>
<th>203</th>
<th>1927</th>
<th>4869</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total used : 3442
Computing Frequency of Use of Code Points

Initial Data Set : 1312
Top 10 Most Popular Words

TC
- 的, 2774
- 人, 1005
- 在, 975
- 一, 964
- 是, 960
- 不, 951
- 中, 896
- 有, 883
- 大, 776
- 台, 718

SC
- 日, 20942
- 月, 20315
- 人, 4430
- 国, 3754
- 中, 3521
- 被, 2791
- 称, 2340
- 南, 2152
- 地, 2226

JP
- 日, 822
- 年, 496
- 国, 393
- 会, 345
- 月, 325
- 人, 325
- 大, 319
- 市, 253
- 本, 250
- 中, 250
- 京, 251
Top 20 : Chinese Frequency of Use > Japanese Frequency of Use

Generated Data Set : 939
Top 20 : Chinese Frequency of Use < Japanese Frequency of Use

Generated Data Set : 363
Chinese Frequency of Use = Japanese Frequency of Use

Generated Data Set : 10

Frequency of Use %

- 8FCE
- 7D20
- 675F
- 541B
- 846C
- 79E9
- 82BD
- 96C0
- 5857
- 5353

C-Freq vs J-Freq
Frequency of Use Reassembly

C / J Usage Overlap Data Set : 1312

Freq J > C : 363
Freq C > J : 939

R_c

J = C
10

Problem Domain (Unsolved Overlap) : 10
Data Processing & Computation Recap

- >20K Han Code Points
- 6181 CJK Overlap
- 1312 Usage Overlap

Filtering Process

LOGIC Design

Frequency of Use Computation

10 Code Points

Methodology Review
CJK Coordination
Re-Sampling & Computation
Statistical Justification

Problem domain was effectively reduced
Future Work

Overlap range redefine
Expand (?) Std Dev.

Require intensive CKJ coordination & deliberation

Mean = 0.034465
S.D. = 0.158477

Chinese Frequency of Use Minus Japanese Frequency of Use

R_j

R_c
Re-consider Language Tag

Policy

TLD registries → IANA/Verisign provisioning

Internet query

root server operators publication

root servers

DNS resolvers

Language tag support

c

j

k

Sources of Language Tag

•RFC 2860: The name space of language tags is administered by IANA
•ISO Standard 639:
  •when a language has both an IANA-registered tag and a tag derived from an ISO registered code, one MUST use the ISO tag.
•Maintenance Agency: International Information Centre for Terminology (Austria)
**Perfection Syndrome**

“Engineering isn't about perfect solutions; it's about doing the best you can with limited resources.” Randy Pausch

---

Thank You

**Question?**

---

Kenny Huang, Ph.D.
huangksh@gmail.com