The Design & Development of Pan-CJK Fonts

Dr. Ken Lunde
Senior Computer Scientist
Adobe Systems Incorporated
lunde@adobe.com

Adobe

What Is A Pan-CJK Font?

- A Pan-CJK font includes glyphs suitable for multiple CJK locales
  - China, Taiwan, Hong Kong, Japan, and Korea are the five most important CJK locales
  - "Han Unification" necessitates multiple glyphs for many CJK Unified Ideograph code points
- A Pan-CJK font is Unicode-based
  - No other character set or encoding in common use today can claim adequate CJK support
  - Unicode has become the preferred method for representing text in digital form
- A Pan-CJK font represents an incredible amount of work—time & effort
- How does a Pan-CJK font differ from a Pan-Chinese font?
  - There are three primary Chinese-speaking locales: China, Taiwan, and Hong Kong
  - Some simplified/traditional distinctions have been unified
  - Some distinctions have not been unified
    - These distinctions are handled via separate code points
  - A Pan-Chinese font can be treated as a first step toward developing a Pan-CJK font

Pan-CJK Font Origins & Goals

- Single-locale CJK fonts can be fully-functional with one glyph per code point
  - This is easily demonstrated by today's single-locale CJK fonts
  - Some single-locale CJK fonts still require multiple glyphs for some code points
    - For the purpose of supporting single-locale variant forms
- Multiple-locale CJK fonts require more than one glyph for some code points
  - CJK Unified Ideograph code points are the obvious target and concern
  - Punctuation and other characters may require locale-specific forms

Pan-CJK Font Advantages

- Typeface design consistency across multiple locales
  - Weight
  - Style
  - Width
  - Relative size
  - Hinting, which influences rendering at smaller sizes and at lower resolutions
  - Other design factors
- Smaller overall footprint
  - A large number of glyphs are shared by two or more locales
  - Subroutinization benefits
    - Applies to OpenType/CFF fonts
- Single font file
- Streamlined testing

CJK Unified Ideographs: URO Versus Extensions

- Premise: CJK Unified Ideograph code points require multiple glyphs
  - Some code points require only one glyph—many are single-source code points
  - Some require more than one glyph—these are multiple-source code points
- Single- versus multiple-source code points
  - Single-source code points generally require only one glyph
  - Multiple-source code points have the *potential* to require more than one glyph
    - Example: U+4E00 (一) has six sources, but clearly requires only one glyph
- URO (Unified Repertoire & Ordering)
  - 20,902 + 22 (Unicode 4.1) + 8 (Unicode 5.1) + 8 (Unicode 5.2) = 20,940 code points
- Extensions
  - Extensions A (6,582), B (42,711), C (4,149), and D (222) exist
  - The higher the Extension, the greater the percentage of single-source code points
  - The higher the Extension, the lower the percentage of multiple-source code points

CJK Unified Ideographs: URO Versus Extensions (cont'd)

(Number of Sources)	1	2	3	4	5	6	7
URO (20,940)	9%	7%	11%	18%	32%	22%	1%
Extension A (6,582)	9%	27%	41%	20%	3%	>0%	>0%
Extension B (42,711)	45%	41%	14%	1%	>0%		
Extension C (4,149)	91%	8%	>0%	>0%			
Extension D (222)	98%	2%					

CJK Unified Ideographs: URO Versus Extensions (cont'd)

- Significantly more "work" is required for URO code points
  - The URO has a high percentage of multiple-source code points
  - Remember that multiple-source code points have the "potential" to require multiple glyphs
    - Not all multiple-source code points require multiple glyphs
- The higher the Extension, the less "work" that is required
  - Higher Extensions have a higher percentage of single-source code points
- Code-point/glyph-count ratios
  - The URO and Extension A require roughly a 50% increase in glyphs over code points
  - Approximately 30K glyphs are necessary to cover the 20,940 URO code points
  - Approximately 10K glyphs are necessary to cover the 6,582 Extension A code points

Locale-specific Glyph Issues

- Different glyphs for the same locale
  - Multiple-column CJK Unified Ideograph code charts versus current source glyphs
  - Some source glyphs have changed over time
    - JIS X 0213:2004 (Japan) is a good example
- Handling CJK Unified Ideographs without sources for specific locales
  - For those specific to a single locale, it is appropriate to ignore
    - Simplified Chinese is a good example
    - U+8BED (语) is tied to a single-locale, specifically Simplified Chinese
  - For the remainder, it becomes a policy issue
    - Extrapolate or ignore

Locale-specific Glyph Issues—Specific Examples

- Source glyphs that changed over time: U+8FBB
  - Original Japanese source glyph:  (JIS X 0208-1990 36-52)
  - Current Japanese source glyph:  (JIS X 0213:2004 1-36-52)
- Multiple-source CJK Unified Ideographs that require only one glyph: U+4E00
  - All sources:
- Two glyphs serve more than two code points: U+5668, U+FA38 & U+20F96
  - U+5668 glyphs
    -  for Japan, and  for all other sources
  - U+FA38 glyph (*ignoring that the distinction that is meant to be preserve cannot be preserved*)
    -  for Japan
  - U+20F96 glyph
    -  for Taiwan

Pan-CJK Font Implementation Details

- TrueType Collection—via separate font instances
  - Pro: 'locl' GSUB feature support is not necessary; no need to choose a default locale
  - Con: Multiple font instances in application font menus
    - *Can be considered a Pro in some uses or contexts*
- OpenType—via 'locl' GSUB feature
  - Pro: Single font instance in application font menus
    - *Can be considered a Con in some uses or contexts*
  - Con: 'locl' GSUB feature support is necessary; must choose a default locale
- Dealing with the 64K glyph barrier
  - Depends on the extent to which CJK Unified Ideograph blocks are covered
  - This is a clear concern when supporting all of Extension B
    - 20,940 (URO) + 6,582 (Extension A) + 42,711 (Extension B) = 70,233

Implementing Pan-CJK Fonts: OpenType

- Use the "Adobe-Identity-0" ROS
  - ROS corresponds to /Registry = "Adobe"; /Ordering = "Identity"; and /Supplement = 0
  - A dynamic, locale-unspecific special-purpose glyph set
- Use the 'locl' (Localized Forms) GSUB feature
  - One locale must necessarily serve as the default
    - Simplified Chinese is suitable due to GB 18030's broad coverage—URO + Extension A
  - The remaining locales are supported via substitutions defined in the 'locl' GSUB feature
    - Language and script tags must be specified
    - Simplified Chinese = ZHS/hani
    - Traditional Chinese = ZHT/hani (Taiwan) and ZHH/hani (Hong Kong)
    - Japanese = JAN/kana
    - Korean = KOR/hang
- Fully-functional prototype fonts have been built

Other Pan-CJK Font Implementations

- TrueType Collection (TTC)
  - Single font file with multiple font instances
    - Appropriate when the font instances can share a significant number of glyphs
  - Each supported locale has its own font instance
  - Separate font "instances" can share common glyphs
    - Application font menus advertise multiple font instances, one for each locale
  - The 'locl' GSUB feature is not necessary
  - Two iPhone fonts, STHeiti-Light.ttc and STHeiti-Medium.ttc, are Pan-CJK TTC fonts
    - Also included in Mac OS X 10.6, but without the Japanese and Korean font instances
      - These are Pan-Chinese fonts

Other Pan-CJK Font Implementations (cont'd)

- OpenType Collection (OTC)
  - *The best of both worlds?*
  - One font instance can use the 'locl' GSUB feature for handling locale-specific glyphs
    - A locale-independent font instance
  - Additional font instances can be tied to specific locales
    - These locale-specific font instances can still use the 'locl' GSUB feature
  - The use of CFF provides a file size advantage

Other Pan-CJK Font Implementations (cont'd)

- Composite Font
  - A Composite Font is a "recipe" that references one or more Component Fonts
  - A Composite Font that can specify fonts by language/script can serve as a Pan-CJK font
  - A Composite Font can be used to overcome or work around the 64K glyph barrier
    - A Composite Font is necessary when dealing with Extension B in its entirety

Other Pan-CJK Font Implementations (cont'd)

- "Pan-CJK" IVD (Ideographic Variation Database) Collection
  - Registered IVSes (Ideographic Variation Sequences) correspond to locale-specific glyphs
    - Allows locale-specific glyph distinctions to be represented in "plain text"
  - Single glyph, multiple sources—U+4E00 (一)
    - 4E00 E01*xx*; Pan-CJK; 4E00-G
    - 4E00 E01*xx*; Pan-CJK; 4E00-T
    - 4E00 E01*xx*; Pan-CJK; 4E00-J
    - 4E00 E01*xx*; Pan-CJK; 4E00-K
  - Multiple glyphs, multiple sources, some shared across locales—U+9AA8 (骨 & 骨)
    - 9AA8 E01*xx*; Pan-CJK; 9AA8-G
    - 9AA8 E01*yy*; Pan-CJK; 9AA8-T
    - 9AA8 E01*yy*; Pan-CJK; 9AA8-J
    - 9AA8 E01*yy*; Pan-CJK; 9AA8-K

Other Pan-CJK Font Implementations (cont'd)

- "Pan-CJK" IVD (Ideographic Variation Database) Collection (cont'd)
  - Single glyph, single source—U+8BED (语)
    - 8BED E01*xx*; Pan-CJK; 8BED-G
  - Serves as a "blueprint" for developing Pan-CJK fonts

Pan-CJK Font Support in OSes & Applications

- OpenType: InDesign CS3 and greater supports the 'locl' GSUB feature
  - Can be specified in character and paragraph tags
- TTC: Mac OS X and Windows can generally handle such fonts
  - Applications enumerate fonts differently, so extensive application testing is required
  - Most TTCs to date have been single-locale
  - Multiple-locale, specifically Pan-CJK, TTCs are relatively new
- OTC
  - Still at the experimentation stage
  - Support in most environments except for Windows
- Composite Fonts
  - Somewhat application-specific
  - Composite Font Standard (CFS) is a forthcoming ISO standard for a Composite Font format
    - Also supports Fallback Fonts

Unicode Coverage Issues

- Which CJK Unified Ideographs should be included?
- Minimal coverage
  - IICore—9,810 CJK Unified Ideographs
    - 9,706 URO, 42 Extension A, and 62 Extension B
- Intermediate coverage
  - Common standards—GB 18030, Hong Kong SCS-2008, JIS X 0213:2004 and KS X 1001:2004
  - Equivalent to URO, Extension A, partial Extension B, and one Extension C code point
    - GB 18030 requires all URO and Extension A code points, plus six in Extension B
    - Hong Kong SCS-2008 requires 1,712 Extension B code points, plus one in Extension C
    - JIS X 0213:2004 requires 303 Extension B code points
- Maximum coverage
  - All of them
  - This obviously breaks the 64K glyph barrier that is inherent in today's font formats

Locale-specific Considerations

- Hangul
  - Specific to Korean
  - 11,172 code points
- Kana
  - Specific to Japanese, but included in standards of China, Taiwan, Hong Kong, and Korea
  - Accounts for 70% of Japanese text, so the typeface design must be very good
  - Requires vertical variants for the small versions and for the long vowel mark
- Vertical variants for punctuation and kana
  - Some vertical variants are locale-specific

Pan-CJK Font Prototype Details

- A "proof of concept" OpenType font
- Makes use of the 'locl' GSUB feature
- 44,000 glyphs
  - Covers 29,925 CJK Unified Ideograph code points
    - URO + Extension A + partial Extension B—close to intermediate coverage
  - Supplied by Changzhou SinoType
- The default locale is Simplified Chinese
  - 11,267 'locl' GSUB feature substitutions for Traditional Chinese
  - 8,106 'locl' GSUB feature substitutions for Japanese
  - 5,312 'locl' GSUB feature substitutions for Korean
- Its glyphs have not been extensively checked for locale appropriateness
- An IICore subset version includes 15,770 glyphs—minimal coverage

Demo

- Adobe InDesign + OpenType Pan-CJK font prototype
  - Specifying locale via paragraph tags
  - Specifying locale via character tags
    - Overrides the locale specified by the paragraph tag on a per-character basis

Future Predictions

- Today's Pan-CJK fonts require multiple glyphs for many code points
  - One cannot argue this point due to locale-specific conventions that transcend typeface design
- In the future, cross-cultural unification efforts are possible
  - Unicode may serve as the catalyst
  - The Web is making the world smaller, and cross-cultural interaction is ever-increasing
  - This is not likely during the current generation, but perhaps within 25 years
- This can be considered *genuine* Han Unification!

Further Reading & Resources

- *CJKV Information Processing*, Second Edition (O'Reilly Media, 2009)

  http://oreilly.com/catalog/9780596514471/

- OpenType Specification

  http://www.microsoft.com/typography/otspec/

- The Unicode Consortium

  http://www.unicode.org/

Adobe