Development of a Pan-CJK Font
What is a Pan-CJK Font?

- Pan (greek: πᾶς) means "all" or "involving all members" of a group
- Pan-CJK means a Unicode based font which supports different countries which use Han ideographic characters: China, Hong Kong, Taiwan, Japan and Korea
- The origin for the necessity of a Pan-CJK font is the Unicode “Han Unification” which uses one common Unicode for different glyphs
- Unicode is the main encoding today for text and also for any kind of information exchange. It includes all local character sets and is a globally accepted standard.
- Supporting different CJK locales means a font with multiple glyphs for one codepoint
Why do we need Pan CJK Fonts?

- Some applications/devices use only one font
- Global data interchange is using Unicode:
  - Using local fonts with different glyphs for the same Unicodes might result in a strange text display.
  - Displaying a Japanese SMS on a Chinese mobile phone
  - Displaying the content of your MP3 player in a car audio system from another locale
  - Display of names in navigation systems/maps
- Having Fonts with different glyphs on the same codepoint is an anachronistic technology (last century!)
Pan European Fonts are easy to create

- You need three scripts: Latin, Greek, Cyrillic
- Nearly all latin codepoints are unique, forms like Scedilla and Scommaacent have different Unicodes:
  - Š uni015E Ū uni0218
- For the Cyrillic script we have some different forms for Serbia and Bulgaria:
  - Ђ ђ Ѣ ѣ ѓ ж ӣ ې լ Ը Ԯ ԯ ժ ʓ ƃ Ɔ ۼ Ӧ Ӗ Ӂ ӂ ӷ ДЉ Ԭ Ӗ Ӂ ӂ ӷ
- “locl” Feature for Romania, Moldavia, Serbia, Bulgaria

That’s all!
Pan CJK Fonts are hard to create

- For Han Unification many sources have been used (Dictionaries, national Standards...)
- For one code point you might have up to seven different sources and therefore up to 7 different glyph shapes
  Some glyphs have multiple sources but a unique shape:
  for example uni4E00: 一
- Some codepoints have many different shapes:
  uni9AAA: 骘 骘 骘
- There is no rule which code points require multiple glyphs, you have to find out by comparing local fonts and consulting the Unicode standard
- Some differences are quite small
- Even different font styles like Song and Hei might have a different number of alternative glyphs for the same code point
- Punctuation for different locales might require different glyphs as well
Han Ideographs in Unicode 6.1

- Unicode defines a total of **74,617** CJK Unified Ideographs
- CJK Unified Ideographs (4E00–9FFF) **20,941** (most have 5 sources)
- Extension A (3400–4DBF) **6,582** (most have 3 sources)
- Extension B (20000–2A6DF) **42,711** (most have 1 source)
- Extension C (2A700–2B73F) **4,149** (most have 1 source)
- Extension D (2B740–2B81F) **222** (most have 1 source)

Ken Lunde (Adobe) has counted the number of sources for each Unicode area.
In all sources you have about **208,900** glyphs
<table>
<thead>
<tr>
<th>Unicode version</th>
<th>Addition</th>
<th>Plane</th>
<th>Characters added</th>
<th>Total Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>CJK Unified Ideographs</td>
<td>Basic Multilingual Plane (BMP)</td>
<td>20,902</td>
<td>20,914</td>
</tr>
<tr>
<td></td>
<td>CJK Compatibility Ideographs</td>
<td>BMP</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>3.0</td>
<td>CJK Unified Ideographs Extension A</td>
<td>BMP</td>
<td>6,582</td>
<td>27,496</td>
</tr>
<tr>
<td>3.1</td>
<td>CJK Unified Ideographs Extension B</td>
<td>Supplementary Ideographic Plane (SIP)</td>
<td>42,711</td>
<td>70,207</td>
</tr>
<tr>
<td>4.1</td>
<td>CJK Unified Ideographs: Ideographs from HKSCS-2004 and GB 18030-2000 not in ISO 10646</td>
<td>BMP</td>
<td>22</td>
<td>70,229</td>
</tr>
<tr>
<td>5.1</td>
<td>CJK Unified Ideographs: Ideographs from Adobe Japan and disunification of U+4039</td>
<td>BMP</td>
<td>8</td>
<td>70,237</td>
</tr>
<tr>
<td>5.2</td>
<td>CJK Unified Ideographs Extension C</td>
<td>SIP</td>
<td>4,149</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8 other characters from ARIB #47, #95, #93 and HKSCS</td>
<td>BMP or SIP</td>
<td>8</td>
<td>74,394</td>
</tr>
<tr>
<td>6.0</td>
<td>CJK Unified Ideographs Extension D</td>
<td>SIP</td>
<td>222</td>
<td>74,616</td>
</tr>
<tr>
<td>6.1</td>
<td>1 character corresponding to Adobe-Japan1-6 CID+20156</td>
<td>BMP</td>
<td>1</td>
<td>74,617</td>
</tr>
</tbody>
</table>
4 Sources, 3 different glyph shapes

5 Sources, 1 glyph
Production of Pan CJK Font

- Starting point: Chinese GB2312-80 and Big-5 (around year 2000)
- Part of Nimbus Sans Global
- Data Format and production system: IKARUS®
  - Format suitable for 4-Byte glyph numbers, no limit
  - Production system can handle composite fonts
  - Easy Conversion to TTC, OTF, TTF
- Step by step expansion:
  - GB18030, Korean Hangul, HKSCS
  - Taiwan alternate glyphs, Hongkong alternates
  - Japanese alternate glyphs
- 2012 Status: > 65000 glyphs
Advantages of Pan CJK Font

- One data base file: easy to manage
- Consistent Design (weight, stem width, style)
- Many glyphs shared for different locales
- Easy to create single locale fonts
- Use of glyph components (elements, radicals, bodyparts) for the design of stylistic alternate (locale specific) glyphs
- Data compression possible
  - Composite glyphs in TTF/TTC
  - Subroutinization in OTF
IKARUS Font Production
CJK Composite Glyphs (SMF)
Future IKARUS Development

- SMF will be developed further to include all locale glyph variants
- This will simplify the design and the production process
- Allow to produce the font as a composite font (Reduce file size)
Technical Implementation of Pan CJK Font

- First Option: TTC (TrueType Collection File)
  - Pro: One font file, no features (locl) necessary
  - Con: Same limit as a single font, multiple names in applications

- Second Option: TTF/OTF (with GSUB features: locl, ss01...)
  - Pro: One font file
  - Con: Feature support in application necessary
  - One locale chosen as default (for example GB18030)
    - Simplified Chinese: ZHS/hani
    - Traditional Chinese: ZHT/hani (Taiwan), ZHH/hani (Hong Kong)
    - Japanese: JAN/kana
    - Korean: KOR/hang
Third Option: IVS (Ideographic Variation Selectors)

- An IVS is an additional Unicode code point which can be used to identify locale specific glyphs
- IVS is supported in OTF/TTF
- Unicode range E0100-E01EF
- Example: uni9AAA

9AAA E0100 鲙 Simplified Chinese

9AAA E0101 𦴖 Traditional Chinese (Taiwan alternates)

9AAA E0102 𦴖 Japanese
IVS (Ideographic Variation Selectors) continued

- Kind of Han De-Unification
- Allows the exact identification of local glyph variants in Unicode based text
- Can be used with the OTF ccmp or liga feature
- IVS should be registered with the Unicode consortium
- Ideally: Create national standard for IVS for all CJK locales and register unique variation selectors for each locale
  Simplified Chinese should serve as default (GB18030 has the largest glyph repertoire)
Technical limitations

- The 64 K (2 byte) barrier in OTF/TTF
  - Limits the number of glyphs to 65535
  - No possibility to cover all CJK Ideographs
  - Even worse for composites: Each component counts as a glyph
  - TTC doesn’t help, same limit as TTF for the tables

- Status of OTF feature (locl...) support
  - Implementation of feature support not consistent

- Existence of language specific Applications (InDesign)
  - Customers in Germany mostly use the standard ID version
  - No support for Japanese features, no support for Arabic, no support for complex scripts
Nimbus Sans Global as a Pan-CJK Font

- GB 18030 (CJK + Extension A + 6 Extension B)
- Hong Kong SCS-2008 (1,712 Extension B, 1 Extension C)
- JIS X 0213:2004 (303 Extension B)
- KS X 1001:2004 (4888 Hanja CJK + Extension A, 11102 Hangul)

Local Glyph variants

- 8504 Japanese glyphs alternates (locl, ss01)
- 8641 Traditional Chinese glyph alternates (locl, ss03) for Taiwan, Hongkong

Total glyph number: 65513 (The end is near!)
(58) Single Substitution

[Example: replace 'uni5542' with 'uni56a2.png']

Dr. Jürgen Willrodt     URW++ Design & Development GmbH
Future Developments

- Pan CJK Fonts require many glyph variants today
- To deal with Pan CJK Fonts today:
  - Adobe & Microsoft et al.
    - Please remove the 64K barrier and define OTF 2.0
  - Adobe: Please release global versions of your Creative Suite
  - All software companies and developers working with fonts: Please implement support for all OTF feature consistently with the same functionality!
- Maybe in the future a unification of glyph design across cultures might be possible driven by the Web and global information interchange!