CJK Integration Algorithm (v0.2)

Precondition

- Each of CJK Generation Panels (GPs) generates LGR for each language TLD before integration.
  - CJK GPs pick up ideographic variants (if any) for Han characters from domain name usage perspective.
  - CJK GPs don't elaborate ideographic variants for Han characters from linguistic perspective.
- CJK GPs agree on the mechanism (steps) to integrate and extract each language (script) LGR.

Step 1: Each CJK GP generates its own LGR (hereinafter, LGR-1)

- LGR-1 generation process is left to each CJK GP.
- LGR-1 format must follow XML schema for LGR.
  < https://datatracker.ietf.org/doc/draft-davies-idntables/>
- Each code point must have individual <char> element (i.e. don't use <range> element).
- Each <char> element in LGR-1 must have reflexive mapping as <var> element (i.e. each code point must have explicit variant type/subtype).

Step 2: CJK GPs collectively generate a merged table of each LGR-1 (hereinafter, LGR-M)

1. Each CJK GP checks if WLE of each LGR-1 has any conflicts. If any, CJK GPs solve the conflicts and back to Step 1.
2. Extract every <char> element tagged "sc:Hani" from each LGR-1.
3. For each extracted <char> element, check the existence of another <char> element with the same code point ("cp" value), and if existent, merge them into one element. At this time, "type" attribute of <var> element must be removed.
4. After the check was finished, record every merged <char> elements to LGR-M.

- Repertoire of LGR-M is the union of all sc:Hani in each CJK LGR-1.
- Variants of each <char> elements in LGR-M is the union of all variants defined for the code point in each CJK LGR-1.
- LGR-M does not have following information.
  - Language tag.
  - Variant type/subtype attribute of <var> elements.
  - WLE (Whole Label Evaluation rules).

Step 3: Each CJK GP extract its original repertoire with integrated variants from LGR-M.

1. For each <char> element in its LGR-1 (hereinafter, <char>::LGR-1), extract <char> element of the same code point ("cp" value) from LGR-M (hereinafter, <char>::LGR-M).
   1.1. For each <var> element in <char>::LGR-M (hereinafter, <var>::LGR-M), compare with <var> elements in <char>::LGR-1 (hereinafter, <var>::LGR-1).
      1.1.1. If <var>::LGR-M has the same code point ("cp" value) with <var>::LGR-1, then copy "type" attribute of <var>::LGR-1 to <var>::LGR-M.
      1.1.2. Otherwise, set "type" attribute with value "blocked" to <var>::LGR-M.
      1.1.3. If code point ("cp" value) of <var>::LGR-M does not match with any code points of <char> elements in LGR-1, record the code point to Out-of-Repertoire list (hereinafter, OoR-list).
   1.2. Record <char>::LGR-M to Integrated-Repertoire list (hereinafter, IR-list).

Step 4: Each CJK GP add "Out of Repertoire" code points for symmetry.

1. For each code point in OoR-list (hereinafter, cp::OoR-list, extract <char> element of the same code point ("cp" value) from LGR-M (hereinafter, <char>::LGR-M).
   1.1. For each <var> element in <char>::LGR-M (hereinafter, <var>::LGR-M), compare cp::OoR-list and code point ("cp" value) of <var>::LGR-M.
   1.2. If the two code points are the same, add "type" attribute with value "out-of-repertoire-var" to <var>::LGR-M.
   1.3. Otherwise, add "type" attribute with value "blocked" to <var>::LGR-M.
   1.4. Record <char>::LGR-M to IR-list.

Step 5: Each CJK GP merge WLE in LGR-1 into one.

1. Each GP extract <rules> element from each LGR-1 and merge them into one WLE (generate integrated <rules> element, hereinafter, <rules>::LGR-M).
2. Each GP add following rule to <rules>::LGR-M for handling “out-of-repertoire-var” variant type.
   <action disp=”invalid” any-variant=”out-of-repertoire-var” />

Step 6: Each CJK GP generates integrated LGR (hereinafter, LGR-2).

1. Each GP extract preambles from its LGR-1 (hereinafter, preamble::LGR-1).
2. Each GP extract all <char> elements with “tag” value other than “sc:Hani” and record to IR-list.
3. Each GP merge preamble::LGR-1, IR-list and <rules>::LGR-M into LGR-2.

- In other words, this step replaces body of <data> element and <rules> element of LGR-1 to IR-list and <rules>::LGR-M respectively.