- ⋆ a list of compositional formulas of glyphs;
- ⋆ a catalogue of ‘factors’, i.e. basic building blocks;
- ⋆ a way of ordering factors (using strokeorders);
- ⋆ an algorithm to derive XXXXXXXXXX

**** KWIC ****

version 1

占	占	
占戈	战	<
占乂	卤	<
占※戌口	鹹	<
占※灸	鹼	<
占灬	点	<
貝|占	貼	<
巾|占	帖	<
贝|占	贴	<
黑|占	點	<
钅|占	钻	<
亻|占	佔	<
魚|占	鮎	<
广|占	店	<
立|占	站	<
米|占	粘	<
氵|占	沾	<
氵|占※	滷	<
臣𠂉|占※皿	鹽	<
禾人氷|占	黏	<

version 2

占	占	
貝|占	貼	>
巾|占	帖	>
贝|占	贴	>
黑|占	點	>
钅|占	钻	>
亻|占	佔	>
魚|占	鮎	>
广|占	店	>
立|占	站	>
米|占	粘	>
氵|占	沾	>
禾人氷|占	黏	>
占戈	战	
占乂	卤	
氵|占※	滷	
占※戌口	鹹	
臣𠂉|占※皿	鹽	
占※灸	鹼	
占灬	点	

version 3

占	占
禾人氷|占	黏
貝|占	貼
巾|占	帖
贝|占	贴
黑|占	點
钅|占	钻
亻|占	佔
魚|占	鮎
广|占	店
立|占	站
米|占	粘
氵|占	沾
占戈	战
占乂	卤
氵|占※	滷
占※戌口	鹹
臣𠂉|占※皿	鹽
占※灸	鹼
占灬	点

version 1

氷	氷	
天|氷	泰	
彐|氷	录	
彐|氷刂	剥	<
𠂤|氷刂	剝	<
尸|氷罒勹虫	屬	<
尸|氷牛	犀	<
尸|氷牛辶	遲	<
石彐|氷	碌	<
日共|氷	暴	<
金彐|氷	録	<
金𠂤|氷	錄	<
礻彐|氷	禄	<
礻𠂤|氷	祿	<
纟彐|氷	绿	<
糸彐|氷	緑	<
糸𠂤|氷	綠	<
禾人|氷占	黏	<
艹月𦍌|氷	藤	<
日日共|氷	曝	<
禾勿人|氷	黎	<
月木人|氷	膝	<
火日共|氷	爆	<
氵木人|氷	漆	<

version 2

氷	氷	
天|氷	泰	
彐|氷	录	>
石彐|氷	碌	>
日共|氷	暴	>
金彐|氷	録	>
金𠂤|氷	錄	>
礻彐|氷	禄	>
礻𠂤|氷	祿	>
纟彐|氷	绿	>
糸彐|氷	緑	>
糸𠂤|氷	綠	>
艹月𦍌|氷	藤	>
日日共|氷	曝	>
禾勿人|氷	黎	>
月木人|氷	膝	>
火日共|氷	爆	>
氵木人|氷	漆	>
氵日共|氷	瀑	>
禾人|氷占	黏	
彐|氷刂	剥	
𠂤|氷刂	剝	
尸|氷罒勹虫	屬	
尸|氷牛	犀	

version 3

氷	氷
天|氷	泰
日共|氷	暴
日日共|氷	曝
火日共|氷	爆
氵日共|氷	瀑
月木人|氷	膝
氵木人|氷	漆
禾勿人|氷	黎
艹月𦍌|氷	藤
彐|氷	录
石彐|氷	碌
金彐|氷	録
礻彐|氷	禄
纟彐|氷	绿
糸彐|氷	緑
金𠂤|氷	錄
礻𠂤|氷	祿
糸𠂤|氷	綠
禾人|氷占	黏
彐|氷刂	剥
𠂤|氷刂	剝
尸|氷罒勹虫	屬
尸|氷牛	犀

version 1

version 2

version 3

True to the original ideas of the KWIC principle, each glyph appears as many times in the index as it has factors. Within the 15,000 or so most common characters, the maximum number of factors per glyph is 6, while the average is a little under 3.

|木缶木冂鬯彡　鬱
木|缶木冂鬯彡　鬱
木缶|木冂鬯彡　鬱
木缶木|冂鬯彡　鬱
木缶木冂|鬯彡　鬱
木缶木冂鬯|彡　鬱

|尸氷皿勹虫　屬
尸|氷皿勹虫　屬
尸氷|皿勹虫　屬
尸氷皿|勹虫　屬
尸氷皿勹|虫　屬

** Formulas **

The compositional character formulas represent the core of the Jizura project. I became aware of the 漢字データベース project sometime around 2007 or 2008; at the time, being able to download thousands upon thousands of formulas under the GPL license was what really jumpstarted this.

The formula language used here is based on the Ideographic Description Language (IDL) as proposed by the Unicode Consortium,[1] with a few extensions; compared to other approaches,[2] IDL has the advantage of being at the right level of abstraction for our purposes and being both human-readable and syntactically straightforward.

```
u-cjk/4e00 一 ●
u-cjk/4e01 丁 ◓一亅
u-cjk/4e02 丂 ◓一㇂
u-cjk/4e03 七 ⊘一乚
u-cjk/4e04 丄 ◓丨一
u-cjk/4e05 丅 ◓一丨
u-cjk/4e06 丆 ◓一ノ
u-cjk/4e07 万 ◓一勹
u-cjk/4e08 丈 ⊘ナ㇏
u-cjk/4e09 三 ◓一二
u-cjk/4e0a 上 ◓卜一
u-cjk/4e0b 下 ◓一卜
u-cjk/4e0c 丌 ◓一川
u-cjk/4e0d 不 ◐𠂇卜
u-cjk/4e0e 与 ◑㇂一
u-cjk/4e0e 与 ◑◐㇂一一
u-cjk/4e0f 丏 ◑◐丁冂厶
u-cjk/4e10 丐 ◓下㇂
u-cjk/4e11 丑 ◑冂土
u-cjk/4e12 丒 ◓刃一
u-cjk/4e13 专 ⊘二◓㇆丶

[...]

u-cjk/4f9c 俜 ◐亻舟
u-cjk/4f9d 依 ◐亻衣
u-cjk/4f9e 伽 ◐亻如
u-cjk/4f9f 侈 ◐亻存
u-cjk/4fa0 侠 ◐亻夹
u-cjk/4fa1 侀 ◐亻丣
u-cjk/4fa2 侅 ◐亻再
u-cjk/4fa3 侶 ◐亻吕
u-cjk/4fa4 侤 ◐亻考
u-cjk/4fa5 侥 ◐亻尧
u-cjk/4fa6 侦 ◐亻贞
u-cjk/4fa7 侧 ◐亻则
u-cjk/4fa8 侨 ◐亻乔
u-cjk/4fa9 侩 ◐亻会
u-cjk/4faa 侪 ◐亻齐
u-cjk/4fab 侫 ◐亻妄
u-cjk/4fac 侬 ◐亻农
u-cjk/4fad 侭 ◐亻尽
u-cjk/4fae 侮 ◐亻每
u-cjk/4faf 侯 ◐亻◓コ矢
u-cjk/4fb0 侰 ◐亻君
u-cjk/4fb1 侱 ◐亻呈
u-cjk/4fb2 侲 ◐亻辰

[...]

u-cjk/513d 儽 ◐亻纍
u-cjk/513e 儾 ◐亻囊
u-cjk/513f 儿 ◐ノ乚
u-cjk/5140 兀 ◓一儿
u-cjk/5141 允 ◓厶儿
u-cjk/5142 先 〓
u-cjk/5143 元 ◓一兀
u-cjk/5144 兄 ◓口儿

[...]

u-cjk/53ad 厭 ◔厂猒
u-cjk/53ae 厮 ◔厂斯
u-cjk/53af 歷 ◔厤㉿心
u-cjk/53af 歷 ◔厤&cdp#x8962;
u-cjk/53b0 厰 ◔厂敞
u-cjk/53b1 厱 ◔厂僉
u-cjk/53b2 厲 ◔厂萬
u-cjk/53b3 厳 ◔◓ゝ厂敢
u-cjk/53b4 厴 ◔厭甲
u-cjk/53b5 厵 ◓原◐原原
u-cjk/53b6 厶 ◐厶丶
u-cjk/53b7 厷 ◔ナム

[...]

u-cjk-xb/20194 覕 ◐享兒
u-cjk-xb/20195 覢 ◐幷夜
u-cjk-xb/20196 覣 (◓亠◐方氏合)
u-cjk-xb/20197 覤 ◓亦合
u-cjk-xb/20198 覮 (◓产八口一仪)
u-cjk-xb/20199 覰 ◐亭夜
u-cjk-xb/2019a 覺 (◓亠眙死)
u-cjk-xb/2019b 覾 ◓亠皾
u-cjk-xb/2019c 覿 ◓亦妻
u-cjk-xb/2019d 覾 (◓亠◐日日八人戊)
u-cjk-xb/2019e 覾 ◐享亶
u-cjk-xb/2019f 覾 (◓亠臾彐豕)
u-cjk-xb/201a0 覾 (◓亠⊘且同百灬)
u-cjk-xb/201a1 覾 (◓亠⊘且同然)
u-cjk-xb/201a2 ㅅ ▿
u-cjk-xb/201a3 天 ◓一人
u-cjk-xb/201a4 介 ◓人亅
u-cjk-xb/201a5 从 ⊘人ノ
u-cjk-xb/201a6 仚 ◔乚◓人一
u-cjk-xb/201a7 仏 ◐亻人
u-cjk-xb/201a8 仔 ◐亻了
u-cjk-xb/201a9 仉 ◐亻凡

[...]
```

[1] Unicode Specification V6.0, Ch. 12 http://www.unicode.org/versions/Unicode6.0.0/ch12.pdf [2] https://en.wikipedia.org/wiki/Chinese_character_description_languages

```
u-cjk-xd/2b81b 齷 ◐齒星
u-cjk-xd/2b81c 齶 ◐齿兒
u-cjk-xd/2b81d 鼇 ◓敞亀

u-cjk-sym/3005 々 ▿
u-cjk-sym/3006 〆 ●
u-cjk-sym/3007 〇 ●
u-cjk-sym/3021 〡 ▿
u-cjk-sym/3022 〢 ◐丨丨
u-cjk-sym/3023 〣 ◐〢丨
u-cjk-sym/3024 〤 ▿
u-cjk-sym/3025 〥 ●
u-cjk-sym/3026 〦 ▿
u-cjk-sym/3027 〧 ◓亠一
u-cjk-sym/3027 〧 ◓丶二
u-cjk-sym/3028 〨 ◓亠二
u-cjk-sym/3028 〨 ◓丶三
u-cjk-sym/3029 〩 ◐ノ又
u-cjk-sym/3038 〸 ▿
u-cjk-sym/3039 〹 ▿
u-cjk-sym/303a 〺 ▿
u-cjk-sym/303b 〻 ●
u-cjk-sym/303d 〽 ●
```

```
u-cjk-rad1/2f00 一 ▿
u-cjk-rad1/2f01 | ▿
u-cjk-rad1/2f02 丶 ▿

[...]

jzr/e100 申 ◓ノ甲
jzr/e100 申 ㊀白ノ
jzr/e100 申 ◓ノ由
jzr/e101 鬼 ◓ノ兎
jzr/e102 艮 ◐目丶

[...]

hzk1/b3d0 &hzk1#xb3d0; ◔厂◐◓土方夂
hzk1/d044 &hzk1#xd044; ◐亻◓隹乃
hzk1/d143 &hzk1#xd143; ◐亻◓廾◔勹用
hzk1/d2ea &hzk1#xd2ea; ◐亻黄

[...]
```

**** Strokeorders (札字五筆法) ****

u-cjk/4e00 一 <1>
jzr/e167 𠂇 <1>
u-cjk/4e8c 二 <11>
u-cjk-xb/2011f 二 <11>
u-cjk-xb/20120 二 <11>
u-cjk/4e09 三 <111>
u-cjk/4e96 三 <1111>
jzr/e15a 丰 <11112>
u-cjk/8a00 言 <1111251>
u-cjk-xb/229d4 戴 <1111342511534>
u-cjk/5f0e 弐 <111154>
u-cjk/5f10 弐 <111154>
u-cjk/4e30 丰 <1112>
jzr/e13f 丰 <1112>
jzr/e14d 丰 <1112>
jzr/e225 丼 <11121112>
u-cjk-xb/234f5 棻 <111211121234>
u-cjk-xb/27bee 彗 <111211121251431>
u-cjk-xb/277f7 靖 <11121112125351>
u-cjk-xa/3ece 爽 <11121112134>
u-cjk-xb/28af9 鑿 <1112111213434112431>
u-cjk-xb/24acd 鑿 <11121112134343112431>
u-cjk-xb/24ad2 鑿 <11121112134345412121>
u-cjk-xb/24ad2 鑿 <11121112134345412121>
u-cjk-xb/21689 褻 <1112111213445434>

**** Factors ****

```
1 一 1              一 1                              一
1 一 1              一 1                              一v1                ㇀
1 一 1              一 1                              一v2   jzr-e167    𠂇

1 一 11             二 11                             二
1 一 11             二 11                             亖
1 一 11             二 11                             二
1 一 11             二 111                            三
1 一 11             二 1111                           𠀎
1 一 11             二 1111111111111111    jzr-e24d   𠀏
1 一 11             二 11112               jzr-e15a   丰
1 一 11             二 11112225211         jzr-e2b3   靑
1 一 11             二 1112                jzr-e14d   𠂉
1 一 11             二 1112                jzr-e13f   王
1 一 11             二 1112                           丰
1 一 11             二 1112                           丰x2h  jzr-e225   丱

[...]

1 一 12  jzr-e21a   亅 12                  jzr-e21a   亅
1 一 12  jzr-e21a   亅 125111152                       𠀋

1 一 12  jzr-e1a6   𠂉 12                  jzr-e1a6   𠂉
1 一 12  jzr-e1a6   𠂉 1215                            弓

1 一 12             丁 12                             丁
1 一 12             丁 12                             丁v1                丅
1 一 12             丁 121                            工
1 一 12             丁 121                            工x4q               𢆶
# 1 一 12            丁 1212                 jzr-e2d6   𠘨
1 一 12             丁 12121                          正
1 一 12             丁 12121                          正xlr               五
1 一 12             丁 12122111            jzr-e21c   耳

[...]

2 | 25              冂 25                             冂
2 | 25              冂 25                             冂v1   jzr-e2e3   冂
2 | 25              冂 25                             冂x2c               冏
2 | 25              冂 251                            囧
2 | 25              冂 251                 jzr-e352   回
2 | 25              冂 251                 jzr-e365   囧
2 | 25              冂 251                            口
2 | 25              冂 251                            口x2c               回
2 | 25              冂 251                            口x2h               叩
2 | 25              冂 251                            口x2v               吕
2 | 25              冂 251                            口x3h               品
2 | 25              冂 251                            口x3p               品
2 | 25              冂 251                            口x4q               𠙵
2 | 25              冂 251                 jzr-e2a9   弓
2 | 25              冂 251                 jzr-e2aa   冃
2 | 25              冂 251                 jzr-e2dd   冄
2 | 25              冂 2511                           冃
2 | 25              冂 2511                           冃
2 | 25              冂 2511                           冄
2 | 25              冂 2511                           日
2 | 25              冂 2511                           日x2h               昭
2 | 25              冂 2511                           日x2v               昌
2 | 25              冂 2511                           日x3p               晶
```

```
2 | 25            冂 2511                       日x4q           閶
2 | 25            冂 2511                       曰

[...]


3 丿 31           𠂉 31                         𠂉
3 丿 31           𠂉 31115                      𠂉
3 丿 31           𠂉 3112                       午
3 丿 31           𠂉 3112                       牛
3 丿 31           𠂉 3112                       牛x2v           隼
3 丿 31           𠂉 3112                       牛x4q           犇
3 丿 31           𠂉 31121                      生
3 丿 31           𠂉 31121                      生x2h           甡
3 丿 31           𠂉 311212                     年
3 丿 31           𠂉 3112121                    缶
3 丿 31           𠂉 31122221          jzr-e18d 無
3 丿 31           𠂉 311234                     朱
3 丿 31           𠂉 31125             jzr-e353 用
3 丿 31           𠂉 311252            jzr-e175 韋
3 丿 31           𠂉 311252                     缶
3 丿 31           𠂉 31134                      失
3 丿 31           𠂉 31134                      矢
3 丿 31           𠂉 3115                       气
3 丿 31           𠂉 312               jzr-e13d 𠂉
3 丿 31           𠂉 312               jzr-e13e 𠂉
3 丿 31           𠂉 312               jzr-e13d 𠂉x2h           竹
3 丿 31           𠂉 312               jzr-e13e 𠂉x3p           笏
3 丿 31           𠂉 312               jzr-e13e 𠂉x6p           篇
3 丿 31           𠂉 312               jzr-e13e 𠂉x8q           篇
3 丿 31           𠂉 3121                       生
3 丿 31           𠂉 3121                       牜
3 丿 31           𠂉 31211                      乍

[...]

4 丶 41           亠 41                         亠
4 丶 41           亠 4111251                    言
4 丶 41           亠 4111251                    言x2h           誩
4 丶 41           亠 4111251                    言x2v           喜
4 丶 41           亠 4111251                    言x4q           囍
4 丶 41           亠 41121                      主
4 丶 41           亠 412234            jzr-e11a 亦
4 丶 41           亠 4124                       卞
4 丶 41           亠 4125125251                 高
4 丶 41           亠 413                        广

[...]


5(12)             𠃌 5                 𠃌 5                          𠃌
5(12)             𠃌 51  jzr-e346      𠃍 51              jzr-e346  𠃍
5(12)             𠃌 51  jzr-e346      𠃍 5121                      𠃎
5(12)             𠃌 51  jzr-e350      𠃊 51              jzr-e350  𠃊
5(12)             𠃌 51  jzr-e350      𠃊 5121121211      jzr-e31e  𢆶
5(12)             𠃌 51  jzr-e1c2      𠃋 51              jzr-e1c2  𠃋
5(12)             𠃌 51  jzr-e1c2      𠃋 5122111         jzr-e21b  𠄎
5(12)             𠃌 51  jzr-e1c2      𠃋 5134                      央

5(12)             𠃌 51  jzr-e341      𠃐 51              jzr-e341  𠃐
5(12)             𠃌 51  jzr-e341      𠃐 511                       ⇉
5(12)             𠃌 51  jzr-e341      𠃐 511                       ⇉x2v  jzr-e26d  彗
```

```
5(12)              ⺄ 51  jzr-e341  ⺕ 511112                  聿
5(12)              ⺄ 51  jzr-e341  ⺕ 51112                   聿
5(12)              ⺄ 51  jzr-e341  ⺕ 511121         jzr-e186  肄
5(12)              ⺄ 51  jzr-e341  ⺕ 5112           jzr-e114  尹
5(12)              ⺄ 51  jzr-e341  ⺕ 5112                    尹
5(12)              ⺄ 51  jzr-e341  ⺕ 51121                   聿
5(12)              ⺄ 51  jzr-e341  ⺕ 5112234        jzr-e115  隶
5(12)              ⺄ 51  jzr-e341  ⺕ 51123134                隶

[...]

5(34)              ㇛ 5            ㇛ 5                       ㇛
5(34)              ㇛ 53  jzr-e1cb  㚔 53             jzr-e1cb  㚔
5(34)              ㇛ 53  jzr-e1cb  㚔 531                     女
5(34)              ㇛ 53  jzr-e1cb  㚔 531                     女x2h          妓
5(34)              ㇛ 53  jzr-e1cb  㚔 531                     女x2v          奚
5(34)              ㇛ 53  jzr-e1cb  㚔 531                     女x3p          姦
5(34)              ㇛ 55            ㇛ 55                      ㇛
5(34)              ㇛ 55            ㇛ 555                     ㇜
5(34)              ㇛ 55            ㇛ 555                     ㇜x6q          ㇜
5(34)              ㇛ 55            ㇛ 5551                    ㇞
5(37)              ㇒ 5             ㇒ 5                       ㇒
5(47)              ㇋ 5             ㇋ 5                       ㇋
5(47)              ㇈ 5             ㇈ 5                       ㇈
5(47)              ㇈ 53  jzr-e27c  ㇀ 53             jzr-e27c  ㇀
5(47)              ㇈ 53  jzr-e27c  ㇀ 534            jzr-e27a  ㇂
5(47)              ㇃ 5             ㇃ 5                       ㇃
5(8)               〓 5             〓 5                       〓
# ideographic space:
5(8)                 5               5
5(8)               ⿴ 5             ⿴ 5                       ⿴
5(8)               ⿰ 5             ⿰ 5                       ⿰
5(8)               ⿲ 5             ⿲ 5                       ⿲
5(8)               ⿳ 5             ⿳ 5                       ⿳
5(8)               § 5             § 5                     §
5(8)               ◯ 5             ◯ 5                       ◯
```

(over 4,000 lines)

垂重熏
厲黳
焦羔
灸灸炙
杰釆
畢重焦煮黑黑魚君鳥黑爲厲熏墨鳳巢峲嶚裹黳
体体伾
侀倕
牀牴
侀侀抓扒
坋扮扮挐
村材
荧荣荣荥
荬荧
茉茉
包包
晋晋普寉晉普晋竟韺普普普韺晋普晉韺晋
晋晉
丧嗇嗇嗇
尚亩亘卣茴
茵商冏商
胄宵寉
痟疳
音吝音育章竞重童竟音
林粖
淌誜
湆湇
葐茴
扷找我牧
我莪戕義蓑戕襄蓑褰簑
宗宗
辰辰辰晨辰
更曳
臭臭
赴赴
垔堊
圇圇
豳豘豳
函
畑器器器器器器器腳
奐罕罣垔垉罟罟罝帚器墾墾
蠹

**** Figural Themes ****

龘龘
另另冎足疋豆至戾
足⻊克克兜兜
訇宫宫
圩打廾扯扛圩扯任玨玷扦圩托牡
妅妣
江汢
吖吐
抎抎
址扯
地扡
垚圪
屯圪
屯垔
扒抨
圩扦
扙圾圾扙
圬扢
圩扜扦
京啻音阜臯亶嘼臯亩臯章京堯壑亩冓禀嘼
戛啅帛高壟鬳桌嘾嗣奠嘾嚎豪堯孛蠹蠹
高嘾皋壟槀壟褒贂奠亹堯鬳鬊豪蓀蓀膏
亭堯毫壼韋章鬊豪堯膏鬳壼蠢膏高毫槀
亳鬳亮嚞孛乞戶戶扃亨享

爾爾
電雷電雷
麻麻
眾衆
阣阣
耋耆鴦鴦鴦鴦
耆耆
耂耆耂
耋耋
斷斷斷
匱匱
滿滿
褻褻
王王玊壬壬
开开
勻刁习勻勺勾句
零零
毯屐屧氅罷疐疑縩
筆筆

**** Shape Identity Mappings (SIMs) ****

(22,000 lines)

To keep u-cjk-xb/2011e 二 'archaic form of 下 上' as distinct from u-cjk/4e8c 二 'two', which has the exact same shape, may be justified by the《康熙字典》, which emulates the《說文解字》in this respect; however, doing so would be incompatible with our 'shapes-only' policy. I do not know what motivated the IRG to encode u-cjk-xb/201a2 𠂉'(人-like, somewhat flattened shape that appears as top component)' as distinct from u-cjk/4eba 人 'man; human'; anyway, u-cjk-xb/201a2 𠂉does not seem to appear in《康熙》and is likewise incompatible with the 'shapes-only' policy.

u-cjk-xb/2011f 二 'archaic form of 下' is in fact (if minimally) distinguishable from both u-cjk-xb/20120 二'archaic from of 二 "two" and u-cjk/4e8c 二 'two', so we keep these codepoints separate—except when writing out character formulas, where we always use u-cjk/4e8c 二 in an attempt to prevent becoming overly specific.

u-cjk/4eba 人 u-cjk-rad1/2f08 人
u-cjk/4eba 人 u-cjk-xb/201a2 𠂉
u-cjk/4e8c 二 u-cjk-rad1/2f06 二
u-cjk/4e8c 二 u-cjk-xb/2011e 二
u-cjk/4e8c 二 u-cjk-xb/2011f 二!components
u-cjk/4e8c 二 u-cjk-xb/20120 二!components

The amount of duplication in Unicode is certainly surprising to the layman; some of this is due to Unicode's 'roundtrip encoding compatibility with legacy encodings', some is deliberate, and some is due to oversight.

As far as CJK Ideographs are concerned, the two biggest single sources for duplicate codepoints are Compatibility Codepoints and the (I believe misguided) attempt to treat well-known 'radicals' (字 典 部 首) and some characters explicitly labeled as 'symbols' apart from the 'ideographs proper'. To do so, the Unicode Blocks u-cjk-rad1/2f00 ... u-cjk-rad1/2fdf *Kangxi Radicals*, u-cjk-rad2/2e80 ... u-cjk-rad2/2eff *CJK Radicals Supplement* and u-cjk-sym/3000 ... u-cjk-sym/303f *CJK Symbols and Punctuation* have been set up; the latter contains, inter alia, u-cjk-sym/3007 〇 'zero' (to me undoubtely a 'character', just as 一, 二, 三, 萬 are characters, not symbols), duplicated as u-cjk-strk/31e3 ㆆ '(circular stroke, used for some ideographs created in Korea)'.

u-cjk-xb/2967f 𩙿 u-cjk-rad2/2ede 𩙿
u-cjk/98df 食 u-cjk-rad1/2fb7 食
u-cjk/98df 食 u-cjk-rad2/2edd 食
u-cjk/98e0 飠 u-cjk-rad2/2edf 飠
u-cjk/9963 饣 u-cjk-rad2/2ee0 饣
u-cjk/98e0 飠 u-cjk-xb/2967f 𩙿!components
u-cjk/91d1 金 u-cjk-cmpi1/f90a 金
u-cjk/91d1 金 u-cjk-rad1/2fa6 金
u-cjk/91d1 金 u-cjk-rad2/2ed2 釒!components
u-cjk/8fb6 辶 u-cjk-cmpi1/fa66 辶
u-cjk/8fb6 辶 u-cjk-rad2/2ecc 辶
u-cjk/8fb6 辶 u-cjk-rad2/2ecd 辶!components
u-cjk/8fb6 辶 u-cjk-rad2/2ece 辶!components
u-cjk/5ef4 廴 u-cjk-rad1/2f35 廴
u-cjk-sym/3023 卅 jzr-fig/e177 卅
u-cjk/4e28 丨 u-cjk-sym/3021 丨
u-cjk/4e28 丨 u-cjk-rad1/2f01 丨
u-cjk/4e28 丨 u-cjk-strk/31d1 丨
u-cjk/4e85 亅 u-cjk-rad1/2f05 亅
u-cjk/4e85 亅 u-cjk-strk/31da 亅
u-cjk/4e28 丨 u-cjk/4e85 亅!components/search
u-cjk/4ea0 亠 u-cjk-rad1/2f07 亠
u-cjk/4ea0 亠 u-cjk-sym/3026 亠
u-cjk/5341 十 u-cjk-rad1/2f17 十
u-cjk/5341 十 u-cjk-sym/3038 十
u-cjk/535d 卄 u-cjk-cmpi1/fa5d 卄
u-cjk/535d 卄 u-cjk-rad2/2ec0 卄
u-cjk-rad2/2ebf 艹 u-cjk-cmpi1/fa5e 艹
u-cjk/5344 艹 u-cjk/8279 艹
u-cjk/5344 艹 u-cjk-rad2/2ebe 艹
u-cjk/5344 艹 u-cjk-sym/3039 艹
u-cjk-xb/25ad7 𥫗 u-cjk-rad2/2eae 𥫗
u-cjk-xb/2099d 卓 u-cjk/9fba 卓
u-cjk-xd/2b740 丩 jzr/e14f NO USE
u-cjk-xc/2b1e6 艹 jzr/e182 NO USE
u-cjk/6075 恵 u-cjk-cmpi1/fa6b 恵
u-cjk/8218 舘 u-cjk-cmpi1/fa6d 舘
u-cjk/6ed1 滑 u-cjk-cmpi1/f904 滑
u-cjk/5951 契 u-cjk-cmpi1/f909 契
u-cjk/61f6 懶 u-cjk-cmpi1/f90d 懶
u-cjk/7669 癩 u-cjk-cmpi1/f90e 癩
u-cjk/908f 邏 u-cjk-cmpi1/f913 邏
u-cjk/4e82 亂 u-cjk-cmpi1/f91b 亂
u-cjk/6feb 濫 u-cjk-cmpi1/f922 濫
u-cjk/85cd 藍 u-cjk-cmpi1/f923 藍
u-cjk/8964 襤 u-cjk-cmpi1/f924 襤
u-cjk/5eca 廊 u-cjk-cmpi1/f928 廊
u-cjk-xb/20628 八 u-cjk-rad2/2e87 八
u-cjk-xb/206a4 々 u-cjk-sym/3005 々
u-cjk-xb/208de 匍 u-cjk-cmpi2/2f9dd 匍
u-cjk-xb/21018 噯 u-cjk-xb/2103c 噯
u-cjk-xb/21d0b 嵓 u-cjk-cmpi2/2f8f8 嵓
u-cjk-xb/21f37 嶲 u-cjk-xb/2439a 嶲
u-cjk-xb/22331 舁 u-cjk-cmpi2/2f891 舁
u-cjk-xb/232ad 䎱 u-cjk-xb/232ab 䎱
u-cjk-xb/23d1e 涇 u-cjk-cmpi2/2f906 涇
u-cjk-xb/23f41 濛 u-cjk-xb/23f9e 濛
u-cjk-xb/23f5e 濯 u-cjk-cmpi2/2f910 濯
u-cjk-xb/243ab 糞 u-cjk-cmpi2/2f91f 糞
u-cjk-xb/24425 糞 u-cjk-xb/2444b 糞
u-cjk-xb/249bc 豎 u-cjk-xb/249e9 豎
u-cjk-xb/24c36 串 u-cjk-cmpi2/2f935 串
u-cjk-xb/24d14 𠂢 u-cjk-rad2/2eaa 𠂢
u-cjk-xb/24fb8 鰮 u-cjk-cmpi2/2f93c 鰮
u-cjk-xb/25249 瞭 u-cjk-cmpi1/fad5 瞭
u-cjk/9091 邑 u-cjk-rad1/2fa2 邑
u-cjk/9149 酉 u-cjk-rad1/2fa3 酉

漢字データベース

```
jzr/e1ef 馬 cdp/896a &cdp#x896a;
jzr/e1f0 甘 cdp/88c8 &cdp#x88c8;
jzr/e1f1 帀 cdp/85fd &cdp#x85fd;
jzr/e1f8 𠕇 cdp/8569 &cdp#x8569;
u-cjk/5e80 庀 c2/215a &c2#x215a;
u-cjk/68b5 梵 c1/5b31 &c1#x5b31;
u-cjk/904b 運 c1/672a &c1#x672a;
u-cjk/9db3 鶳 c2/6d34 &c2#x6d34;
u-cjk-xb/26ddf 𦷟 cb/15d5 &cb#x15d5;
u-cjk/5e97 店 c1/4d33 &c1#x4d33;
u-cjk-xa/387b 㡻 c3/2866 &c3#x2866;
u-cjk/7385 玅 c2/2c28 &c2#x2c28;
u-cjk/9d19 鴙 c2/5c39 &c2#x5c39;
u-cjk/6b63 正 c1/465f &c1#x465f;
u-cjk/9117 鄗 c1/673a &c1#x673a;
u-cjk/7380 玀 c1/7b66 &c1#x7b66;
```

yadda yadda

**** Dictionary Data ****

The relatively little dictionary data I have comes largely from Unicode's `UniHan.txt` data and Jim Breene's EDICT project[3]. Overall, it is of very mixed quality and would be in need of a thorough overhaul.

```
@glyphs 冰氷
u-cjk/51b0 冰 py:bīng
u-cjk/51b0 冰 ka:ヒョウ
u-cjk/51b0 冰 hi:こおり, ひ, こお･る
u-cjk/51b0 冰 hg:빙
u-cjk/51b0 冰 gloss:ice; ice-cold

@glyphs 氫氢
u-cjk/6c2b 氫 py:qīng
u-cjk/6c2b 氫 gloss:amonia; hydrogen nitride

@glyphs 況况
u-cjk/6cc1 況 py:kuàng
u-cjk/6cc1 況 ka:キョウ
u-cjk/6cc1 況 hi:まし･て, いわ･んや, おもむき
u-cjk/6cc1 況 hg:황
u-cjk/6cc1 況 gloss:condition, situation;
  furthermore

@glyphs 渾浑
u-cjk/6d51 浑 py:hún
u-cjk/6e3e 渾 ka:コン
u-cjk/6e3e 渾 hi:すべ･て, にご･る
u-cjk/6e3e 渾 hg:혼
u-cjk/6e3e 渾 gloss:muddy, turbid; blend,
  merge, mix

@glyphs 游遊
u-cjk/6e38 游 py:yóu
u-cjk/6e38 游 ka:ユウ, リュウ
u-cjk/904a 遊 ka:ユウ, ユ
u-cjk/6e38 游 hi:あそ･(び|ぶ), およ･ぐ
u-cjk/904a 遊 hi:あそ･(ぶ|ばす)
u-cjk/6e38 游 hg:유
u-cjk/6e38 游 gloss:to swim; float, drift;
  wander, roam
```

[3] http://www.edrdg.org/wwwjdic/wwwjdicinf.html#dicfil_tag

**** Variants and Usage ****

This file contains around 8,400 records that connect 'variant characters' and 'regions of usage'.

The regions as used here are `C` for the PRC, `J` for Japan, `T` for Taiwan, `H` for Hong Kong, `M` for Macau, and `K` for North and South Korea; this reflects both how these countries and territories are represented in the Unicode Consortium, and, in particular, in the data presented as a result of the Unicode IICore (International Ideographs Core, 國際表意文字核心, 東アジアの諸国で一般に使用される漢字集合)[4] effort.

The term 'variant character' is understood as an umbrella term for what is variously called 俗字, 古字, 本字, 略字, 異體字, 簡/繁體字, 新字体 and so on in the traditional literature; no effort has been made to differentiate beyond making simple statements like 'glyph A [which is used in regions X...] is a variant of glyph B [which is used in regions Y...]'. In the software, variant relationships are modeled as both symmetric and transitive[5]; further, there is no temporal aspect whatsoever encoded. All of these assumptions make the display and the handling of the data simpler, but they also oversimplify somewhat.

Here are some samples:

```
01 也CJKTHM  艹  辶  芒
02 鷗JKTHm  鸥C  鴎J
03 飢JKTHm  饑KTHm  饥C
04 個JKTHM  箇JKTm  个CJ  𠂆  仝
05 團JKTHM  糰T  团C  団J  糐
06 紆  紓
07 龜KTHM  龟C  亀J  皸  皲  皹  龜  龝  龜  龜
  龜  龜  龝  龝  龜  竈
08 亞JKTHM  亚C  亜J
09 畝JKTHM  亩C  畆  畝  晦  畝  畮
10 偎cTHm  隈cJt  渨
11 台CJKTHM  颱JKTHM  臺KThM  檯TM  枱th  儓  傝
  彙  允  壘  壇  薹  垲  傝  臺  臺  壜  臺  臺
12 元CJKTHM  圓JKTHM  円JK  圆C  圓
```

Line `01` tells us that 艹, 辶and 芒are variants of 也. Of these, only 也 has a 'usagecode'—`CJKTHM` in this case—attached to it from which we can immediately see that the IICore team did not include any of 艹, 辶, 芒in their listing of important characters; in other words, these glyphs are presumably of minor importance for everyday communication. Conversely, `CJKTHM` implies that the glyph 也 is used in all six regions under consideration.

In line `02`, 鸥 is marked as used in the PRC and 鴎 as used in Japan; the other regions use 鷗, as it is marked `JKTHm`. Both 鴎 and 鷗 show up with `J`, the implication being that both are current in Japan, possibly for different usages (I guess 鷗 is used in names).

[4] `https://zh.wikipedia.org/wiki/`國際表意文字核心

[5] that is, when 'A is a variant of B' holds, then 'B is a variant of A' also holds, and when additionally 'B is a variant of C' is true, then 'A is a variant of C' is also true.

**** Frequency and Rankings ****

Using data from the Leeds Corpus[6] for Chinese (PRC) and Japanese character usages, a frequency list compiled by Chih-Hao Tsai[7] and a NodeJS module[8] that promises to perform a reasonable ranking using Bayesian statistics, a little over 15,600 characters mentioned in at least one of the above sources were given a ranking index, starting from #00,001 for the most frequent character, 的, to #15,677 for the least common one, 氖.

It must be said that the procedure as outlined above is probably not very scientific; it does not take the IICore selection of characters into account and neglects potentially useful material such as counts derived from Wikipedia articles, the lists published in connection with language proficiency tests such as the JLPT[9] or the HSK[10]. On the other hand, there is no way to make any meaningful usage statistic without handling huge corpuses and making decisions about which texts to in- and exclude from the counts.

As it stands, our listing does seem to make some intuitive sense, and has been a valuable tool to facilitate the sifting out of tens of thousands of Unicode CJK code points describing characters which virtually never appear in modern written material.

```
00,001 的
00,002 人
00,003 一
00,004 中
00,005 上
00,006 要
00,007 大
00,008 在
00,009 出
00,010 以
00,011 自
00,012 他
00,013 年
00,014 可
00,015 多
00,016 家
00,017 能
00,018 生
00,019 好
00,020 本
00,021 得
00,022 日
00,023 前
00,024 子
00,025 用
00,026 方
[...]
01,008 漂
01,009 爸
01,010 実
01,011 陳
01,012 电
01,013 融
01,014 飲
01,015 架
01,016 歳
01,017 預
01,018 籍
01,019 陣
01,020 关
01,021 憲
01,022 県
01,023 貼
01,024 狗
01,025 続
01,026 妙
01,027 漢
01,028 減
01,029 査
01,030 藏
01,031 搞
01,032 変
01,033 见
01,034 緊
01,035 棒
01,036 発
01,037 楽
[...]
15,664 洚
15,665 檑
15,666 搋
15,667 憨
15,668 艽
15,669 艚
15,670 耖
15,671 痄
15,672 鼢
15,673 糈
15,674 稃
15,675 塌
15,676 螵
15,677 氖
```

[6] http://corpus.leeds.ac.uk/list.html [7] http://technology.chtsai.org/charfreq/ [8] https://github.com/mumme/smart-ranking [9] http://www.jlpt.jp/e/ [10] http://www.chinesetest.cn/index.do