Character indexing

To: UTC, RMG
From: Robin Leroy
Date: 2026-03-05

Background: Index.txt and charindex.html

The UCD contains a data file Index.txt. As documented in UAX #44, this file "is used to maintain the Unicode Character Name Index". The file (and the character name index itself, which is derived from Index.txt with minimal processing) is a manually maintained permuted index of character names and of some other information present in the character names list, such as block names, subheaders, or informal aliases.

See, *e.g.*, the following entries, with sources noted in comments:

```
A WITH ACUTE, LATIN CAPITAL LETTER   00C1
A WITH ACUTE, LATIN SMALL LETTER     00E1
A WITH BREVE, LATIN SMALL LETTER     0103
A WITH CARON, LATIN SMALL LETTER     01CE
...
ABBREVIATION MARK, ARMENIAN          055F
ABBREVIATION MARK, SYRIAC            070F
ABBREVIATION SIGN, DEVANAGARI        0970
Abbreviations, Squared Latin         3371  #  Subheader in block CJK Compatibility.
Aboriginal Syllabics Extended, Unified Canadian
                                     18B0  #  Block name.
Aboriginal Syllabics, Unified Canadian
                                     1400  #  Block name.
...
above, double dot                    0308  #  Informal alias for COMBINING DIAERESIS.
absolute value                       007C  #  Comment on VERTICAL LINE:
                                               • used in pairs to indicate absolute value
...
ARMENIAN ABBREVIATION MARK           055F
```

As part of the Unicode release process, the Release Management Group tracks the updating of Index.txt. This task had traditionally been performed by Joe Becker. The Release Management Group identified as early as 2022 that this index had become unmaintainable and needed to be redesigned; in the meantime, recent updates have been done by Ken Whistler.

Being manually maintained, the index only covers a small subset of the information from the charts: it has merely 6155 entries, including all permutations, whereas the source file for names list annotations has 26 900 entries; no attempt is made at indexing the répertoire of the siniform scripts below the block level. While the use of U+007C | in absolute values is indexed, most comments indicating usage in phonetic notation are not: for instance, there is no index entry for "uvular", even though 9 comments include that word. The choices of permutation are sometimes inconsistent: U+20C1 appears only as RIYAL SIGN, SAUDI (and in particular is nowhere to be found under S), whereas U+0E3F appears under three entries,

- BAHT, THAI CURRENCY SYMBOL
- CURRENCY SYMBOL BAHT, THAI

- THAI CURRENCY SYMBOL BAHT

Further, the index is not consistently updated when names list annotations are corrected. For instance, the annotations for U+2135–U+2138 were corrected based on feedback L2/24-223 ID20240911040714 (see 181-A4), but they are still indexed as

cardinal, first transfinite	2135
cardinal, fourth transfinite	2138
cardinal, second transfinite	2136
cardinal, third transfinite	2137
...	
transfinite cardinal, first	2135
transfinite cardinal, fourth	2138
transfinite cardinal, second	2136
transfinite cardinal, third	2137

Proposal

We propose replacing the current static charindex.html in the Unicode 18.0 time frame by a search tool based on automated indexing of names list annotations and selected properties[1]. This takes advantage of the more closely reviewed technical and editorial work done on the names list and the UCD. The details of the indexing approach and construction of resulting index entries could be refined at any time, independently of the release schedule; the automation makes updating the index for a new release trivial. The indexing tool would be maintained as part of the Unicode tools, leveraging existing tooling for the production and testing of the UCD.

This proposal would make the file Index.txt obsolete; we propose foregoing the update of this file in the Unicode 18.0 cycle (that is, including a copy of Index.txt, Unicode Version 17.0 in Unicode Version 18.0), and removing it from the UCD starting with Unicode Version 19.0.

A preview of this tool can currently[2] be seen at eggrobin.github.io/unicode-annotations/charindex.html. As of this writing, the underlying index has 66 666 entries (not counting permutations[3]), indexed under 37 382 distinct words. The index is part of the page, and the results are computed client-side as the query is typed.

[1] Besides the Block, Name, and Name_Alias properties also surfaced in the character names list, the current draft uses radical-stroke properties, as well as non-property data from CJKRadicals.txt, to index the siniform scripts, based on advice from Ken Lunde and Ken Whistler. The tooling can readily accommodate other properties describing the function or appearance of characters; kEH_Func and kEH_FVal are already incorporated via names list comments, but kEH_Desc, Unihan readings, *vel sim*. could be added to the corpus if it is determined that they are a practical way to find characters.

[2] The preview might be removed after the proposal is reviewed by the UTC. This document includes examples illustrating the behaviour of the current preview. If the proposal is accepted, the search tool will eventually be found at https://www.unicode.org/charts/charindex.html.

[3] A complete list of all permuted entries in the style of Index.txt would have approximately 268 000 entries.

Examples

With this approach, the first few results for a search for "a acute" are

A WITH ACUTE, LATIN CAPITAL LETTER	00C1	Á
A WITH ACUTE, LATIN SMALL LETTER	00E1	á
A WITH BREVE AND ACUTE, LATIN CAPITAL LETTER	1EAE	Ắ
A WITH BREVE AND ACUTE, LATIN SMALL LETTER	1EAF	ắ
A WITH CIRCUMFLEX AND ACUTE, LATIN CAPITAL LETTER	1EA4	Ấ
A WITH CIRCUMFLEX AND ACUTE, LATIN SMALL LETTER	1EA5	ấ

Searching for "absolute" finds

absolute continuity	2AA1	⪡
absolute value, used in pairs to indicate	007C	|

Searching for "cardinal" finds

cardinal	1FA50	□
cardinals. They are used in notations of transfinite		
Hebrew letterlike math symbols. In Letterlike Symbols:	2135–2138	א–ד

Searching for "uvular" finds

uvular ejective stop [q']	A72B	ꝯ
uvular fricative or approximant, voiced	0281	ʁ
uvular fricative, see 0281	A727	ꜧ
uvular implosive, voiced	029B	ʛ
uvular implosive, voiceless	02A0	ʠ
uvular nasal, voiced	0274	ɴ
uvular stop, voiced	0262	ɢ
uvular trill, represents a voiceless	1D29	ᴘ
uvular trill, voiced	0280	ʀ

The results of a search for "badger" illustrate the indexing of the CJK répertoire alongside the scope of the names list:

BADGER	1F9A1	🦡
BADGER, KANGXI RADICAL	2F98	豸
Characters with this radical (153):		
In CJK Unified Ideographs Extension A:	4756–4766	貁–貥
In CJK Unified Ideographs:	8C78–8C9C	豸–貜
In CJK Unified Ideographs Extension B:	27CA0–27D23	豹–貣
In CJK Unified Ideographs Extension C:	2B38A–2B38D	豺–貍
In CJK Unified Ideographs Extension E:	2C948–2C94D	貃–貗
In CJK Unified Ideographs Extension F:	2E665–2E66C	豕–貙
In CJK Compatibility Ideographs Supplement:	2F9D3	
In CJK Unified Ideographs Extension H:	320A4–320A7	貤–貧
In CJK Unified Ideographs Extension J:	32471	𲑱
	33089–33090	𳢉–𳢊

The index also supports radical-stroke values, *e.g.*, a search for “12.5” finds

12.5, CJK radical-stroke		
In CJK Unified Ideographs Extension A:	34B5–34B7	㒵㒶㒷
In CJK Unified Ideographs:	5175	兵
In CJK Unified Ideographs Extension B:	2050A–2050F	𠔊𠔋𠔌𠔍𠔎𠔏
In CJK Unified Ideographs Extension F:	2D047	𭁇
In CJK Unified Ideographs Extension J:	3247F	𲑿
12.5, Jurchen radical-stroke	18ED8–18ED9	□□

cf. the radical-stroke index RSIndex.pdf, p. 14, and, in RSIndex.txt,

```
12.5  U+5175 U+34B5 U+34B6 U+34B7 U+2050A U+2050B U+2050C U+2050D U+2050E
U+2050F U+2D047 U+3247F # 兵㒵㒶㒷𠔊𠔋𠔌𠔍𠔎𠔏𭁇
```

An automated index is bound to contain irrelevant results that no human indexer would list; consider, for instance, in a search for “actually”, besides the possibly relevant

ACTUALLY EQUAL TO, APPROXIMATELY BUT NOT	2246	≆
ACTUALLY EQUAL TO, NEITHER APPROXIMATELY NOR	2247	≇

the more anecdotal

actually used in German dialectology. It is not	AB3E	ꬾ

However, since the index is filtered by search terms, rather than presented as a flat alphabetical list, and since the results are sorted to prioritize concise and structured data such as blocks, subheaders, and character names over the more discursive comments and subheader notices, this seems unlikely to be a major problem in practice.