WG2 N5253
Date: 2024-05-24

Title: Considerations concerning ISO/IEC 10646 7th edition

Source: Michel Suignard, Project editor
Status: Individual Contribution
Distribution: WG2
Reference: SC2/N4890 ISO/IEC 14651 7th edition CD text

Executive Summary:

Since the Unicode Standard and ISO/IEC 10646 editions/amendment are increasingly sharing the same repertoire, we should aim at avoiding work duplication by increasing the number of normative references from ISO/IEC 10646 to Unicode specifications and data set. This would make larger section of ISO/IEC stabilized by not requiring updates as the repertoire is augmented. The goal would be a situation analogous to ISO/IEC 14651 where the draft of the 7th edition is presented as a stabilized document.

History

The considerations described in this document originated from the work done on the latest version of ISO/IEC 14651 – International string ordering and comparison – Seventh Edition, Committee Draft. In that CD, the task of separating the variable content from the stable content has resulted in stabilizing the document content and referencing externally the variable content. This was possible because ISO/IEC 14651 is tightly synchronized with the Unicode Technical Standard UTS #10, Unicode Collation Algorithm. While ISO/IEC 14651 uses its own data named Common Template Table (CTT), it is created through the same process that creates the parallel data structure generated for UTS#10 called the Default Unicode Collation Element Table (DUCET). These CTT and DUCET can be generated for each new Unicode version; therefore, by exposing both tables in a revised version of UTS #10, ISO/IEC 14651 can just refer normatively to the section of UTS#10 describing the CTT data. This allows ISO/IEC 14651 to become a stabilized standard.

In similar fashion, ISO/IEC 10646 and the Unicode Standard are tightly connected. Many elements of the core content of ISO/IEC 10646 are defined by normative references to the Unicode Standard elements, such as the Unicode Bidirectional Algorithm (UAX #9), and the Unicode Normalization Forms (UAX #15). In addition, some data sets are also referenced normatively such as the ‘Age’ property (in Unicode DerivedAge.txt file), and the General Category Property (GC). In the past, this was sometimes challenging because, while the Unicode Standard and ISO/IEC are synchronized in principle, the actual code repertoires were slightly different because the code points were added to each of these standards in different schedules, creating challenges for sharing data elements between the two standards.

Fortunately, the Unicode Standard versions and ISO/IEC 10646 have increasingly shared the same repertoire, recent examples:

- Unicode 13.0 – ISO/IEC 10646: 2020 6th Edition
- Unicode 15.0 – ISO/IEC 10646: 2020 6th Edition augmented with Amendment 1: 2023

And soon, it is expected that Unicode 16.0 will share the same repertoire as ISO/IEC 10646: 2020 6th Edition augmented with Amendment 1: 2023 and Amendment 2: 2025 (expected). Note that because Unicode issues new versions more frequently than ISO/IEC issues new editions/amendments, Unicode 14.0 and 15.1 have no ISO/IEC equivalents.

First step: Stabilizing properties

ISO/IEC 10646 has already some properties specified by normative references to Unicode:

ISO/IEC 10646 Normative references	**Unicode file**
Combining Classes	UnicodeData.txt
General Category	UnicodeData.txt
Bidi Mirrored	UnicodeData.txt
Age	DerivedAge.txt

And other normative references, while not directly referencing Unicode based properties, make use of additional Unicode properties, such as UAX#9 (The Unicode Bidirectional Algorithm), UAX#15 (Unicode Normalization Forms), and UTS#37 (Ideographic Variation Database).

Now considering all existing ISO/IEC 10464 electronic data attachments:

1. Variation selectors and variation sequences:
http://standards.iso.org/iso-iec/10646/ed-6/en/UCSVariants.txt
2. Emoji variation sequences:
http://standards.iso.org/iso-iec/10646/ed-6/en/emoji-variation-sequences.txt
3. Source references for pictographic symbols
http://standards.iso.org/iso-iec/10646/ed-6/en/EmojiSrc.txt
4. Source references for CJK ideographs:
http://standards.iso.org/iso-iec/10646/ed-6/en/CJKSrc.txt
5. Source references for Nüshu characters:
http://standards.iso.org/iso-iec/10646/ed-6/en/NushuSrc.txt
6. Source references for Tangut ideographs:
http://standards.iso.org/iso-iec/10646/ed-6/en/TangutSrc.txt
7. Named UCS Sequence Identifiers:
http://standards.iso.org/iso-iec/10646/ed-6/en/NUSI.txt
8. Collection MOJI-JOHO-KIBAN IDEOGRAPHS-2016:
http://standards.iso.org/iso-iec/10646/ed-6/en/JMJKI-2016.txt
9. Collection MOJI-JOHO-KIBAN IDEOGRAPHS-2018:
http://standards.iso.org/iso-iec/10646/ed-6/en/JMJKI-2018.txt
10. Collection JAPANESE JIS X 0213:2004 IDEOGRAPHS FROM PREVIOUS JIS STANDARDS:
http://standards.iso.org/iso-iec/10646/ed-6/en/JIS-X-0213-FromPrevious.txt
11. Collection JAPANESE CORE KANJI:
http://standards.iso.org/iso-iec/10646/ed-6/en/JapaneseCoreKanji.txt
12. Alphabetically sorted list of character names:
http://standards.iso.org/iso-iec/10646/ed-6/en/Allnames.txt
13. Names of Hangul syllables:
http://standards.iso.org/iso-iec/10646/ed-6/en/HangulSy.txt

Some of these data sets represent collections such as JMJKI-2016.txt, JMJKI-2018.txt, JIS-X-0213-FromPrevious.txt, JapaneseCoreKanji.tx which are immutable. The HangulSy.txt file represents the names of Hangul syllables and is also fixed, because the Hangul Syllables correspond to a fully defined stable set.

The remaining 8 attachments with repertoire dependencies have direct Unicode equivalents:

ISO/IEC 10646 name	Unicode name (in UCD)
UCSVariants.txt	StandardizedVariants.txt
emoji-variation-sequences.txt	emoji/emoji-variation-sequences.txt
EmojiSrc.txt	EmojiSources.txt
CJKSrc.txt	[Unihan.zip] Unihan_IRGSources.txt
NushuSrc.txt	NushuSources.txt
TangutSrc.txt	TangutSources
NUSI.txt	NamedSequences.txt
Allnames.txt	~ extracted/DerivedName.txt

(DerivedName.txt is ordered by code point while Allnames.txt is ordered by character name)

Given this, it could be simpler to just link to the Unicode UCD files, instead of posting parallel files that inherently contain the same content (except for the header file) when the code repertoire is identical.

Next step: Stabilizing clauses

To a degree, this has already happened for parts such as the Bidi Algorithm and the Normalization Forms. The next step could be to simplify some ISO/IEC 10646 clauses that have equivalent and, in many cases, more detailed versions of the same content in the Unicode Standard. For example, Clause 24 'Source references for CJK ideographs' could be replaced by a normative reference to UAX #38 'Unicode Han Database (Unihan)' and refer to the sub section concerning the IRG contribution (kIRGSources). The sub-clause 23.4 Source references for pictographic symbols would be vastly improved by replacing by a link to UAX #51 'Unicode Emoji'. The current text in clause 23.4 is very dated and only corresponds through its link to EmojiSrc.txt to the historical nature of the emoji characters and not to its current status.

The other two sources (Tangut and Nüshu) may require more work on the Unicode side because they don't have actual UAXs related to them and are only briefly described in UAX #44 'Unicode Character Database'. For example, their data format is not formally introduced (unlike CJK sources) but just described as being like Unihan format (as far as the author could determine). However, as it happened for ISO/IEC 14651, it is possible to modify the Unicode version of these documents to make their use as references by an ISO standard more palatable.

Case study: Sources Reference for Egyptian Hieroglyphs

In CDAM2.3 the source references for Egyptian Hieroglyphs were documented in a new text (clause 27) and the corresponding text file: Hieroglyphsources.txt. At the same time, Unicode 16.0 was introducing a new Draft Unicode Standard Annex: UAX #57: Unicode Egyptian Hieroglyph Database https://www.unicode.org/reports/tr57/ which includes the same content in more details. It is preferable to simply link to it as a normative reference to avoid duplication of effort and to improve synchronization between Unicode and ISO/IEC 10646.

Based on this, CDAM2.4 added a normative reference to UAX #57 'Unicode Egyptian Hieroglyph Database (Unikemet)' and created a brief description as follows:

> ### 27. Source references for Egyptian hieroglyphs
>
> An Egyptian hieroglyph is always referenced by one or more source references and ancillary data. The source reference information and the ancillary data establish the character identity for Egyptian hieroglyphs characters. The ancillary data provided by the database define additional information such as a detailed description of the character, various sources, catalogue entries, and function. It also defines properties related to these

hieroglyphs, such as belonging to a Core set, whether they rotate or not, and whether they mirror or not. The information is provided in the Unicode Egyptian Hieroglyph Database (Unikemet) (see Clause 2).

-end of document-