WG2 N5216
SC2 4856
Date: 2023-05-19

Title: Referencing String ordering and comparison from ISO/IEC 10646

Source: Michel Suignard (ISO/IEC Project Editor)
Status: Project Editor contribution
Distribution: SC2, WG2

Summary:

This document proposes to include a normative reference in ISO/IEC 10646 to the Unicode Technical Standard #10: Unicode Collation Algorithm along with a new clause providing a brief description of the algorithm, and an appendix explaining the past synchronization with ISO/IEC 14651. This would allow the continuation of a formal ISO based reference to the String ordering comparison currently provided by ISO/IEC 14651 which is encountering maintenance and update challenges. Both specifications are based on the same data set.

Details

ISO/IEC 14651 is the current International Standard for International string ordering and comparison developed by ISO/IEC. It was maintained in synchronization with the Unicode Collation Algorithm (UCA). Although the presentation and text of the two standards are rather distinct, the approach toward the architecture of multi-level collation weighting and string comparison is closely aligned. In particular, the synchronization between the two standards is built around common data tables which define the default (or tailorable) weights. The UCA adds many additional specifications, implementation guidelines, and test cases, over and above the synchronized weight tables. This relationship between the two standards is similar to that is maintained between the Unicode Standard and ISO/IEC 10646.

Due to the unfortunate passing of the current editor of ISO/IEC 14651 project and the difficulty in establishing a continuing synchronization effort, it may be easier in the future to include by reference in ISO/IEC 10646 a normative reference to the Unicode UTS#10 Unicode Collation Algorithm. Any new update to the repertoire shared by the Unicode Standard and ISO/IEC 10646 is always accompanied by a similar update to UTS #10 which includes a new data set to allow the collation of the updated repertoire. In addition, UTS #10 is benefiting of its tight association with the Unicode Common Locale Data Repository (CLDR) which provides key building blocks for software to support the world's languages, with the largest and most extensive standard repository of locale data available. This should facilitate harmonization of the collation work between Unicode and ISO/IEC communities.

Another area of concern is that the text of ISO/IEC 14651:2020 may not be available on the long term in the list of publicly available standards at https://standards.iso.org/ittf/PubliclyAvailableStandards/. It may be possible to retrieve it on the ISO or IEC Web Store but would require communication with these organizations.

Details about the suggested new text in ISO/IEC 10646 are provided below. However, the text of the new Appendix (V) could be extended to provide a more complete explanation concerning differences between the data sets provided in UTS#10 (DUCET) and ISO/IEC 14651 (CTT).

New normative reference in clause 2 (**Normative References**):

Unicode Technical Standard, UTS #10, *Unicode Collation Algorithm*:
https://www.unicode.org/reports/tr10/tr10-47.html

Clause xx. Collation

Collation is the general term for the process and function of determining the sorting order of characters sequences in a list of such sequences. As a method it can be applied to strings containing characters from the full repertoire specified by this document. It is also applicable to subsets of that repertoire. Collation implementations deal with the complex linguistic conventions for ordering text in specific languages and provide for common customizations based on user preferences.

The starting point of the collation algorithm is the Default UCS (or Unicode) Collation Element Table (DUCET), which is data specifying the default collation order for all characters. Then the Collation algorithm takes an input character sequence and a Collation Element Table, containing mapping data for characters. It produces a sort key, which is an array of unsigned 16-bit integers. Two or more sort keys so produced can then be binary-compared to give the correct comparison between the character sequences for which they were generated.

The detail of the implementation is provided in the Unicode Technical Standard UTS#10 (see Clause 2).

Appendix V
(Informative)

Relationship between ISO/IEC 14651:2020 and Unicode Collation Algorithm

ISO/IEC 14651 (2020) is the last International Standard for International string ordering and comparison developed by ISO/IEC. It was maintained in synchronization with the Unicode Collation Algorithm (UCA). Although the presentation and text of the two standards are rather distinct, the approach toward the architecture of multi-level collation weighting and string comparison is closely aligned. In particular, the synchronization between the two standards is built around the data tables which define the default (or tailorable) weights. The UCA adds many additional specifications, implementation guidelines, and test cases, over and above the synchronized weight tables. This relationship between the two standards is similar to that maintained between the Unicode Standard and ISO/IEC 10646.

For each version of the UCA, the Default Unicode Collation Element Table (DUCET) [Allkeys] is constructed based on the repertoire of the corresponding version of the Unicode Standard. The synchronized version of ISO/IEC 14651 had a Common Template Table (CTT) built for the same repertoire and ordering. The two tables are constructed with a common tool, to guarantee identical default (or tailorable) weight assignments. The CTT for ISO/IEC 14651 was constructed using only symbols, rather than explicit integral weights, and with the Shifted option for variable weighting.

The detailed synchronization points between versions of UCA and published editions (or amendments) of ISO/IEC 14651 are shown in *Table xx*

Table xx. UCA and ISO/IEC 14651

UCA Version	UTS #10 Date	DUCET File Date	ISO/IEC 14651 Reference
15.0.0	2022-08-26	2022-08-09	---
14.0.0	2021-08-27	2021-07-10	---

13.0.0	2020-02-07	2020-01-28	14651:2020 (6th ed.)
12.1.0	2019-04-26	2019-04-01	---
12.0.0	2019-03-04	2019-01-25	---
11.0.0	2018-05-10	2018-02-10	---
10.0.0	2017-05-26	2017-04-26	14651:2018 (5th ed.)
9.0.0	2016-05-18	2016-05-16	14651:2016 Amd 1
8.0.0	2015-06-01	2015-02-18	14651:2016 (4th ed.)
7.0.0	2014-05-23	2014-04-07	14651:2011 Amd 2
6.3.0	2013-08-13	2013-05-22	---
6.2.0	2012-08-30	2012-08-14	---
6.1.0	2012-02-01	2011-12-06	14561:2011 Amd 1
6.0.0	2010-10-08	2010-08-26	14561:2011 (3rd ed.)
5.2.0	2009-10-08	2009-09-22	---
5.1.0	2008-03-28	2008-03-04	14561:2007 Amd 1
5.0.0	2006-07-10	2006-07-14	14561:2007 (2nd ed.)
4.1.0	2005-05-05	2005-05-02	14561:2001 Amd 3
4.0.0	2004-01-08	2003-11-01	14561:2001 Amd 2
9.0 (= 3.1.1)	2002-07-16	2002-07-17	14561:2001 Amd 1
8.0 (= 3.0.1)	2001-03-23	2001-03-29	14561:2001
6.0 (= 2.1.9)	2000-08-31	2000-04-18	---
5.0 (= 2.1.9)	1999-11-22	2000-04-18	---

--

References

ISO/IEC 14651: Information technology — International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering:
https://standards.iso.org/ittf/PubliclyAvailableStandards/c079392_ISO_IEC_14651_2020(E).zip

Unicode Technical Standard, UTS #10, *Unicode Collation Algorithm*:
https://www.unicode.org/reports/tr10/tr10-47.html

IEC WEB store:
https://webstore.iec.ch/

ISO WEB Store:
https://www.iso.org/store.html/