Title:	**Unicode Consortium Comments Regarding Maintenance of ISO 12199**
Source:	Unicode Consortium
Source doc reference:	L2/21-154
Author:	Peter Constable and Dr. Kenneth Whistler
Status:	Liaison Contribution
Action:	For consideration by TC 37/SC 2
Date:	August 18, 2021

Unicode has noted the following resolution from the recent TC 37/SC 2 meeting:

> **Resolution 2021-03**
> ISO/TC37/SC2 resolves to establish a WG (Alphabetical ordering, revision in order to work on a minor revision of ISO/12199 Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet after the SR ballot closes, for editorial updates only. …

Also noting that ISO 12199 is currently in systematic review, Unicode would like to comment on the on-going maintenance of ISO 12199 and recommend that it be withdrawn.

Purpose and usefulness of ISO 12199

Stated purpose and scope as success metrics

The stated purpose and scope of ISO 12199 are:

> *"… to have uniform and internationally recognized rules for the alphabetical ordering of terminological and lexicographical data, to make these terminologies more easily accessible for the users… [and to] facilitate the interchange of terminological and lexicographical data."* (Introduction)
>
> *"This International Standard specifies the sequence of characters to be used in the alphabetical ordering of multilingual terminological and lexicographical data (terms, term elements, or words) represented in the Latin alphabet. Character sets of languages represented in the Latin alphabet are taken into account insofar as terminological or lexicographical data have been recorded. Character sets used in internationally standardized transliteration into Latin script are also taken into account."* (Clause 1)

The extent to which this standard is successful, therefore, can be evaluated based on:

- Whether it is used in the implementation of information systems (databases, applications) used for terminological and lexicographic data.

- The extent to which it covers the characters used in orthographies of languages written in Latin script that are of interest for purposes of terminology or lexicography.

Evaluating success of ISO 12199

By both criteria outlined above, we question the usefulness of ISO 12199 as an international standard.

First, in relation to implementations or applications of the standard, we have not found any software implementations that reference ISO 12199, nor have we found any other standards or specifications that reference it.

Moreover, it is noted that ISO 12199 provides a formal and normative specification of its ordering in Annex G by drawing from data included in a different standard, ISO/IEC 14651. This other standard, produced by JTC 1/SC 2, is maintained in synchronization with a standard produced by The Unicode Consortium, Unicode Technical Standard #10, Unicode Collation Algorithm (UTS #10).[1] These specifications provide a default ordering for all characters in Unicode and ISO/IEC 10646, along with mechanisms for language-specific tailoring. In contrast to ISO 12199, which is not widely implemented (if at all), these standards are very widely implemented in software libraries, operating system platforms and database systems.

The only substantial, normative addition in ISO 12199, beyond what is provided in ISO/IEC 14651 and UTS #10, is the specification of word-by-word ordering in Annex A. This is actually a very minor tailoring of the default ordering provided in 14651/UTS #10, assigning a different primary-level weight to SPACE or other word-separating characters.

The second criterion for success pertains to coverage of languages. It is unclear how to determine exactly what languages are to be considered of interest for terminology or lexicography purposes. Some general observations can be made, however:

- Commercial software and content vendors continue to expand the set of languages that they support in content or in language processing features such as spell checking, text prediction, or internationalized domain names.
- Smaller language communities are increasingly engaging in language development efforts, such as language documentation, revitalization efforts and language education.
- Linguists around the world engage in documentation of thousands of modern and historic languages.

These scenarios will involve lexicographic data, if not also terminological data.

Thus, for ISO 12199 to be successful, it really would need to be actively maintained to encompass languages of interest. In 1999, the same year that ISO 12199 was published, Unicode version 3.0 already had 820 Latin characters, though ISO 12199 covers less than 500 (per Annex D). Today, Unicode 13.0 has

[1] For details on the relationship between UTS #10 and ISO/IEC 14651, including history of synchronization, see Annex B of UTS #10: https://www.unicode.org/reports/tr10/#Synch_ISO14651.

1,220 Latin characters. ISO/IEC 14651 and ISO/IEC 10646 are actively maintained by JTC 1/SC 2 and will support all 1,220 Latin characters in their next amendments, along with additional Latin characters that will be in Unicode versions 14.0 and 15.0.

Conclusion from the evaluation

Based on the intended purpose and scope stated in ISO 12199 itself, therefore, ISO/IEC 14651 and UTS #10 clearly are better standards: they encompass the purpose and scope of ISO 12199 but are far more widely implemented and actively maintained. Moreover, ISO 12199 normatively depends on 14651 (though see below on this reference), yet without adding substantial benefit beyond 14651. In addition, both 14651 and UTS #10 are publicly available to any user at no cost.[2]

Therefore, given the availability of these far better standards and no evidence of actual use of ISO 12199 and no substantial additional normative value that it provides*, we suggest that ISO 12199 is not providing significant benefit as an international standard and recommend that it be withdrawn as an IS.*

While the normative content of ISO 12199 does not provide a significant benefit, some might find some of the informative content useful:

- Ordering rules for chemical names (Annex C)
- Character usage for particular languages (Annex D)
- Languages using Latin script (Annex E)
- Tailored ordering for certain languages (Annex F)

Apart from ordering of chemical names, we suggest that Unicode offers better, actively-maintained data for the other types of information as part of the open-source Unicode Common Locale Data Repository (CLDR) project:

- For character usage for particular languages , see the Main Exemplars chart page: By-Type Chart: Main Exemplars (unicode-org.github.io)
- For use of Latin and other script languages by language, see the Scripts and Languages chart: https://unicode-org.github.io/cldr-staging/charts/latest/supplemental/scripts_and_languages.html
- For language specific tailored ordering, see the Collation Charts page: https://unicode-org.github.io/cldr-staging/charts/latest/collation/index.html

If TC 37/SC 2 wishes to continue providing a document with such informative data, it should replace the international standard ISO 12199 with a technical report of more limited scope.

[2] See the ISO/IEC Publicly Available Standards at https://standards.iso.org/ittf/PubliclyAvailableStandards/index.html.

Maintaining ISO 12199 as a normative specification

In its current state, ISO 12199 has issues that undermine its status as a normative specification. As noted above, TC 37/SC 2 is preparing to start a project to revise the standard, but only to make limited editorial changes. That will not correct the technical flaws. If ISO 12199 is to be maintained as an international standard, that can only be worthwhile if it were technically adequate as a normative specification.

The following is a summary of technical issues that would need to be addressed to make it current and technical adequate as a normative specification.

Normative references

The current edition makes normative reference that are obsolete:

> ISO/IEC 10646-1:1993, Information technology — Universal Multiple-Octet Coded Character Set (UCS) — Part 1: Architecture and Basic Multilingual Plane.

The two-part 10646 series was long ago replaced with a consolidated standard, now in its sixth edition: ISO/IEC 10646:2020.[3] JTC 1/SC 2 has recently initiated a project for Amendment 1. Changing this reference would be a simple technical fix.

> ISO/IEC 14651:—, Information technology — International string ordering — Method for comparing character strings and description of a default tailorable ordering.

This reference is undated: it referred to a forthcoming first edition that was still in development. The merits of having normative dependency on an incomplete specification are dubious. Like 10646, 14651 has been actively maintained by JTC 1/SC 2 and is now also in its sixth edition, ISO/IEC 14651:2020, with Amendment 1 initiated to stay in sync with Amendment 1 of 10646. Updating this reference would also be a simple technical fix.

As noted above, though, additional Latin character have been added to Unicode and 10646 / 14651 over the years and may continue to be added in the future. If TC 37/SC 2 wishes to retain ISO 12199 as an international standard, then at a minimum it should plan to *maintain* it with amendments or new editions to remain synchronized with 10646 / 14651 as more Latin characters are encoded.

Annex G

A prose description of the normative process for ordering is given in clauses 4 to 8. While some of this description might be used in a software implementation, it is the formal specification in Annex G that is most likely to be used in implementations of ISO/IEC 14651 and UTS #10.

Annex G has significant flaws, however:[4]

[3] ISO/IEC 10646:2020 is made freely available at https://standards.iso.org/ittf/PubliclyAvailableStandards/index.html.

- There are instances of duplicated and missing symbols in the published table. In principle, these are semantic errors, but the missing symbols would also cause software implementations to fail.
- There are syntactic errors that fail in conformant implementations of ISO/IEC 14651.
- Syntactic conventions are used that were modified in ISO/IEC 14651 after the first edition.

Also, Annex G makes use of a mechanism from 14651 to tailor the default ordering by re-ordering certain sections within the table. On initial review, it is unclear whether these uses in Annex G data have, in fact, the effect it appears was desired. Deeper investigation would be needed to determine this.

In addition, the data in Annex G is organized in a sub-optimal manner that could make it more difficult to integrate into implementations of 14651/UTS #10 and would also result in reduced performance.

A major flaw in Annex G is that it does not take into consideration Unicode Standard Annex #15 Unicode Normalization Forms, which is a normative part of ISO/IEC 10646 and, hence, essential for ISO/IEC 14651.

For example, Unicode and ISO/IEC 10646 allow for two different *but formally equivalent* encoded representations of LATIN SMALL LETTER A WITH ACUTE, “á”: one as a single, precomposed character, U+00E1, and another as a combining-mark sequence, <U+0061 ”a”, U+0301 “◌́”>. While the default collation table of 14651 / UTS #10 takes this into consideration, the data in Annex G of 12199 does not.

The consequence of this is that the result of sorting data using Annex G may be inconsistent, depending on the history of the data and interaction with other processes that might convert it to one Unicode normalization form or another (“NFD”, decomposed, versus “NFC”, composed).

For ISO 12199 to be technically adequate as a normative specification, these issues in Annex G would need to be addressed by preparing *and testing* a significantly revised table.

In our opinion, this would be possible, but would not be worth the significant investment required: What the table in Annex G of ISO 12199 is attempting to specify is already completely covered by the default collation table provided in 14651. Functionally, Annex G is simply a very, very small subset of the default collation table of 14651 and doesn't add anything new in terms of behavioral outcome for ordering of Latin strings.

[4] As mentioned above, ISO 12199 was prepared with reference to a *pre-publication draft* of ISO/IEC 14651. Errors in Annex G of ISO 12199 do not indicate that there are corresponding errors in published editions of ISO/IEC 14651.