L2/18-244

# Title: Unicode Liaison Report to ISO/IEC JTC1/SC2

Date: 2018-6-8
Source: Unicode Consortium
Status: Liaison contribution
Action: For review by SC2 members
Distribution: SC2

The Unicode Consortium is pleased to report on-going progress in development of the Universal Character Set resulting from collaboration with SC2, as well as progress on the Unicode Standard and related standards and technologies.

## Publication of Unicode 11.0

The Unicode Standard, Version 11.0 was just recently published, on June 5, 2018. This version includes the character repertoire of ISO/IEC 10646:2017 (fifth edition), plus Amendment 1, plus the following characters in Amendment 2:

- 46 Georgian Mtavruli capital letters
- 5 urgently-needed CJK unified ideographs (U+9FEB to U+9FEF)
- 66 pictographic symbols used by ICT[1] vendors as emoji

The total count of new characters added (in comparison to Unicode Version 10.0) is 684. These additions include seven newly-encoded scripts.

This release includes all content that comprises the Unicode Standard, including the core text as well as code charts, data files, and annexes that provide detailed implementation specifications such as the Unicode Bidirectional Algorithm.

Certain other technical standards published by Unicode have character data files that are impacted by repertoire additions in new versions of The Unicode Standard. These have also been updated to synchronize their version numbers to 11.0:

- Unicode Technical Standard #10, Unicode Collation Algorithm (analog to ISO/IEC 14651)
- Unicode Technical Standard #39, Unicode Security Mechanisms (addresses spoofing and other security issues related to text processing)
- Unicode Technical Standard #40, Unicode IDNA Compatibility Processing (compatible processing of non-ASCII URLs)
- Unicode Technical Standard #51, Unicode Emoji (emoji-related data and behavior).

In the future, these other standards will remain version synchronized with The Unicode Standard.

[1] Information and communication technology

## Future publication schedule for the Unicode Standard and Unicode Emoji

As has been mentioned in previous liaison reports, Unicode has moved to a yearly publication cycle for new versions of The Unicode Standard. Since 2014 / Unicode Version 7.0, the release date for each version has been in June. Starting in 2019 / Unicode Version 12.0, the release will be moved three months earlier, to March. This is being done to better accommodate vendor product release schedules.

In recent years, ICT vendors have faced significant user pressure for new emoji. To accommodate vendor demands, Unicode has had to release new versions of UTS #51, Unicode Emoji, roughly twice a year. To align with character repertoire in The Unicode Standard and ISO/IEC 10646, new emoji characters would only be added in alternate releases, with other mechanisms used to provide additional emoji in between. Even so, this pace of development for emoji created challenges for Unicode-internal processes as well as for Unicode-SC2 collaboration.

In the meantime, the vendor context has changed, and vendors are now able to work with a once-per-year emoji update. Accordingly, new versions of UTS #51 will from now on be published once per year in synchronization with the annual release of The Unicode Standard.

## Encoding of pictographic symbol characters

Recently, certain SC2 members have expressed concern at Unicode's publication of new pictographic symbol characters while they are still under technical ballot within SC2, and have indicated that they expect there to be willingness to accept additional symbol characters requested by SC2 members. With this in mind, Unicode would like to clarify its position on criteria for encoding of pictographic symbol characters.

It has long been an operational principle for SC2 and for Unicode that proposals for encoding of additional characters should be accompanied by evidence of user-community demand and need for public data interchange — we want to encode characters that have a demonstrated need for encoding and that will, in fact, be used in public data interchange. In most cases, one type of evidence that is expected in order to demonstrate such need is *prior attested usage* in existing publications (whether print or other physical impressions, or digitally).

However, there are valid exceptions to this requirement of demonstrating prior attested usage in widespread publications. The following are examples:

- New currency symbols: When the central bank of a nation decides to adopt a new symbol for the national currency starting at some future date, ICT vendors have a need to support the new symbol in their implementations prior to its widespread usage. In this case, it is institutional commitment to the new character that is important, not prior usage of the new symbol.
- Urgently-needed CJK unified ideographs: Often there is urgent need to support new CJK ideographs that represent new concepts, such as new elements, or to represent special names that need to be recorded in government databases. These are encoded based on the relevant national body indicating anticipated national usage without requiring evidence of prior attested usage in publications.
- Nationally-supported orthographic changes: Occasionally, major institutions within a nation may establish consensus to adopt an orthography change that involves new characters. Some recent

cases of this have been the introduction of an upper/lower case distinction in Georgian and Cherokee scripts. In such cases, it is institutional support for the new characters that is considered sufficient evidence of a need and usage, not prior attested usage.

- The up-coming change Japanese imperial-era change is a special example: institutional commitment for a new era-name ligature character is sufficient evidence, and ICT vendors have a need to support it before it goes into widespread usage.

A common factor in all these examples is that there is a high degree of certainty of future usage.

Another common factor in these examples is that the new characters are needed for key data processes in business, education, government operations, or other such endeavors. This is consistent with the very purpose for which ISO exists and develops standards:

> *ISO creates documents that provide requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose.*[2]
>
> *International Standards* ***make things work****. They give world-class specifications for products, services and systems, to ensure quality, safety and efficiency. They are instrumental in facilitating* ***international trade****.*[3]

In the context of SC2 and Unicode, characters are *not* encoded in ISO/IEC 10646 and Unicode simply for the purpose of recording visual elements that may have been conceived at some point in human cultural history. Rather, then are encoded in the UCS in order to make text processing and public data interchange work in the pursuit of business, educational, governmental, cultural or other endeavors.

With these things in mind, we turn to consider encoding of pictographic symbols, and symbols used by ICT vendors as emoji in particular.

When ICT vendors come to Unicode indicating that they intend to support some emoji, such as a T-Rex emoji, we are given — and require — a clear indication from vendors of intent to support that emoji in their products. In this case, it is the assessment of the Unicode Technical Committee that the requirement of evidence of usage and of need is met, without a requirement to demonstrate prior, attested usage.

Moreover, encoding of the pictographic-symbol character to be used as emoji is found to be consistent with the stated purpose of ISO and Unicode standards: it is done to *make things work* — in this case, to enable communication systems that use emoji in public data interchange to work.

For these reasons, Unicode considers it reasonable to expect that SC2 would be supportive of encoding these characters for which future usage is certain and that are needed for business usage by ICT vendors, in the same way that SC2 supports the encoding of new currency symbols or urgently-needed CJK ideographs.

---

[2] https://www.iso.org/standards.html

[3] https://www.iso.org/about-us.html --- emphasis in the original

Now, when Unicode, or the US national body, presents to SC2 a repertoire of new pictographic symbols that will be used by ICT vendors as emoji, we understand and expect that SC2 members may have comments. Because of the urgency to provide new emoji that vendors have faced in recent years, Unicode has invited WG2 experts with interest in this area to participate in Unicode's processes for evaluating new emoji proposals, which are open to any good-faith contributors, so that input can be provided at the earliest opportunities. That is a standing invitation.

Also, as discussed during the previous SC2 meeting in Hohhot, Unicode agrees to certain procedures to engage WG2 experts for review of characters for new emoji, or any other characters. To wit:

> The UTC should provide review opportunity for new character repertoire, names, and code points before freezing new character repertoire for publication as a new Unicode version.
> ....
> The WG2 email list, wg2@unicode.org, should be used to discuss new character proposals that need to be fast-tracked by the UTC before a ballot opportunity is available. Debbie Anderson and Michael Everson agreed to monitor feedback on this WG2 email list and submit a summary of the feedback to the UTC through the UTC feedback mechanism. Any WG2 expert feedback will be submitted in advance of quarterly UTC meetings.[4]

UTC has been applying these procedures. For instance:

- Notice was sent to the WG2 email list on January 9, 2018 requesting comments on DAM1 and PDAM2, and advising that UTC would need to finalize repertoire for Unicode 11.0 at an up-coming UTC meeting during the week of January 23 – 26.
- Notice was sent to the WG2 email list on March 25 of a one-month review period for comments on the Unicode 11.0 beta.

No feedback for the new emoji in Unicode 11.0 was received from WG2 experts during these two review periods, however. For the future, UTC will continue to be open to and invite feedback from WG2 experts per this process.

Unicode also understands that SC2 members may respond to a new repertoire of pictographic symbols by thinking of yet other symbols that might be encoded. This is welcomed. At the same time, we would expect the same criteria for evaluation to be applied:

- Is there evidence for a text-processing and public-data-interchange need in some business / educational / etc. endeavor?
- Is there evidence of prior, attested usage in text; or a commitment from ICT vendors or other major institutions for future usage?

For example, if an SC2 member were to see a proposed addition for a T-Rex emoji and respond by identifying ICT vendors within their country that were also committed to implementing other dinosaur emoji, then that would make a strong case for encoding of additional pictographic-symbol characters for those other dinosaurs.

[4] SC2/N4965 Summary of Ad Hoc Meeting on SC2/WG2 and UTC Processes

If, however, an SC2 member were to respond by saying that additional dinosaur symbols should be added to make a more-complete set, then we would expect SC2 to require additional evidence to support those additions: that there is a real need for making text processes in public endeavors to work, and that there is a high degree of certainty (perhaps in the form of prior, attested usage) that those symbols will, in fact, be used.

In practice, then, we see pictographic symbols that are potential candidates for encoding as falling into two classes:

- There is the general class, for which the same evidence of need and of usage are required, just as for new orthographic characters. For instance, SC2 has received proposals from SC35 to encode keyboard-related symbols but has yet to approve their encoding due to lack of such evidence.
- Then there is the class of pictographic symbols to be used as emoji: these require the same types of evidence,[5] but such evidence is already implied by the immediate intent of ICT vendors to implement them as emoji.

## Unicode Technical Committee feedback on Shuishu

Unicode experts have reviewed proposals and comment documents for Shuishu script. While it is anticipated that proposals will eventually reach a sufficient level of maturity for encoding, we do not feel that a sufficient level of maturity and consensus on a proposal has yet been attained. This is reflected in recent WG2 documents from Japanese and UK experts:

- WG2/N4942 Comments on Shuishu in PDAM2.2 text: "Shuishu script in PDAM2.2 code chart has several points to be resolved before the standardization"
- WG2/N4946 Additional Comments on Shuishu in PDAM2.2 text
- WG2/N4956 Analysis of Shuishu character repertoire: "We consider that the … academic understanding of the [Shuishu] script is sufficiently mature to allow for the encoding to progress. However, we believe that the repertoire of 486 Shuishu logograms proposed in PDAM 2.2 still has some issues that need to be resolved, and a revised proposal needs to be submitted."

Therefore, we recommend that SC2 postpone encoding of Shuishu to a future edition or amendment, after Amendment 2 of the 5th edition; and we encourage on-going collaboration among Chinese and other experts towards preparation of a mature and sound proposal.

[5] Vendors and UTC also apply additional requirements for new emoji, such as clear, visible distinction at expected text sizes, and expansion of concepts rather than minor distinctions.