Title:	Proposal to maintain UCS repertoire via a maintenance agency
Source:	Peter Constable
Status:	Individual contribution
Action:	For consideration by UTC, INCITS/CS&I
Date:	2023-10-28

Summary

This is a proposal to change the process for maintaining the character repertoire of the ISO/IEC 10646 standard ("UCS"). Up to now, new character repertoire is added to the UCS through amendments and new editions of the standard, requiring national body commenting and balloting through the normal committee, enquiry, and approval stages. The proposal is to stabilize the text of the standard that specifies the architecture for UCS encoding, and to establish a Maintenance Agency that would maintain the UCS repertoire. This could be implemented by JTC1/SC2 in the next edition of the standard.

This is presented for consideration by UTC and INCITS/CS&I, with a proposal that the change described be submitted by the US national body as a proposal to JTC1/SC2, with endorsement from UTC.

Background

The text of the ISO/IEC 10646 standard includes many clauses that specify the architecture of UCS encoding, along with content describing the character repertoire of the UCS. In the 6th edition (2020), code charts documenting the repertoire comprise over 95% of the document. While the code charts evolve with additions to the repertoire, the smaller portion of the content has been stable for some time, without substantive changes in new amendments or editions.

In its early stages of development, while it was important to ensure compatibility with several important national encoding standards, there were many national bodies actively involved in the development of the UCS. Today, only a small number of national bodies remain actively involved (that is, providing expert input in WG2 or IRG), and most of these continue to participate largely from interest in on-going work on CJK ideographs (mostly in the context of IRG).[1]

Moreover, the practice for several years now is that expert contributions for preparing recommendations on new character repertoire largely fall into two categories:

- CJK ideographs

---

[1] Currently, JTC1/SC2 has 24 P-members, but less than half of those have provided experts actively participating in WG2 or IRG. On recent ballots, typically around half of the national bodies abstain, and only a handful of members ever provide comments.

- All other scripts and symbols

Expertise on CJK ideographs is provided mainly by IRG, with significant contributions also provided by UTC experts. In practice, most of the expert input on other scripts and characters has been coming through UTC.[2]

Thus, the situation for some time has been that new repertoire is added to UCS through relatively frequent amendments or editions that follow ISO balloting process through national bodies, yet

- architectural portions of the content have remained stable;

- only a small number of national bodies have been providing expert input in development of the repertoire, mainly via IRG; and

- expert input on non-CJK characters and symbols is increasingly contributed mainly from UTC.[3]

Current SC2 processes don't align well with de facto practice or with needs for on-going development of the UCS repertoire.

The current SC2 process is also limited in agility: new characters must go through an IS-development cycle that takes 18 to 36 months. In contrast, UTC is able to progress urgently needed characters from proposal to publication in a new Unicode version in under 12 months. Recent examples of this include:

- Collaboration with Japan on encoding of hentaigana in Unicode Version 10.0.

- The square ligature for the Japan era name Reiwa in Unicode Version 12.1.

- The addition of China's urgently needed ideographs as CJK Extension I in Unicode Version 15.1.

This difference in agility adds to challenges in synchronization between UCS and The Unicode Standard.

ISO maintenance agencies

It could be argued that maintenance of UCS repertoire was never well suited to normal ISO processes for developing standards. This is seen in the history of frequent amendments throughout its history.[4] The ISO-IEC directives allow for maintenance of a frequently modified standard via a maintenance agency ("MA"):

---

[2] Current work on Seal Script is an exceptional case.

[3] For most scripts that remain to be encoded, the technical issues that need careful consideration pertain to details in implementation such as input, display or other text processing. This is best provided by UTC experts familiar with these aspects of implementation.

[4] In the six years following publication of ISO/IEC 10646-1, there were 31 amendments, 25 of which were for addition of new scripts or characters. Since publication of ISO/IEC 10646:2003, five new editions and 14 amendments have been published, mostly for addition of new repertoire. Most of the time, two or more projects have had to be developed in parallel.

2.11 Maintenance agencies

When a committee has developed a document that will require frequent modification, it may decide that a maintenance agency is required. Rules concerning the designation of maintenance agencies are given in Annex G.[5]

Annex G of the Directives requires that creation of a MA be approved by the ISO Technical Management Board, accompanied by a *rules of procedure* document for the MA that is also approved by the ISO/TMB. Beyond that, the Directives don't impose significant constraints on how a MA must operate.

One example of an ISO coding standard that is maintained by an MA is ISO 3166.

A recent example of the creation of a MA for a coding standard is in relation to ISO 639: A new edition, ISO 639:2023, defines the overall structure and rules for language identifiers, but also establishes a MA for maintenance of the language code. The Terms of Reference of the ISO 639 Maintenance Agency delegates to *language coding agencies* ("LCAs"—InfoTerm, Library of Congress, and SIL) responsibility for publishing the code tables and for gathering and vetting of change requests. A panel of voting members consults with relevant experts and decides on proposals. This panel is comprised of appointees from TC 37/SC 2 and TC 46/SC 4 as well as representatives from the LCAs and from a small number of external stakeholder organizations, such as Unicode. Standards Council of Canada is designated as Secretariat, providing administrative support.

Proposal for UCS

The nature of ISO/IEC 10646 would make it very amenable to maintenance of the repertoire by a MA, and current practice as well as the history of UCS also suggest that this would make sense. Repertoire maintenance by a MA would also improve agility in responding to urgently needed characters, such as CJK Extension I.

A new edition of ISO/IEC 10646 would be prepared that defines the architecture for UCS. It would also establish creation of a MA and specify general rules for maintenance, such as stability requirements. The standard would be significantly restructured, however, with details related to the coded character repertoire—including 2600+ pages of code charts—removed from the text of the standard. A *Terms of Reference* document would be prepared in parallel, defining the composition and processes for the MA.

Given its current role in providing expert input on CJK ideographs, IRG should continue to operate as in the past in preparing recommendations for encoding of new CJK unified ideographs. Similarly, given its current role in providing expert input on CJK ideographs and on all other characters, UTC should also be given a prominent role in the maintenance process.

Also, given that Unicode is already preparing and publishing code charts and data files for the Unicode Standard, and has also contributed the code charts that have gone into past editions of UCS, it would be sensible for Unicode to take responsibility for preparation and publication of code charts and data files.

---

[5] ISO/IEC Directives, Part 1, 2023 Edition.