Title:	**Motivations for establishing a UCS maintenance agency**
Source:	Peter Constable and Robin Leroy (Unicode)
Status:	Expert contribution
Action:	For consideration by SC2 / AHG1
Date:	2025-02-18

0. Abstract

SC2 resolution M29-08 established AHG1 to review the work process for maintaining ISO/IEC 10646 (UCS). This arose from a US national body proposal (SC2/N4888) to create a maintenance agency (MA) for future maintenance of the UCS character repertoire. This document presents reasons why creation of this MA is strongly motivated. First, and most important, it will ensure the long-term viability for on-going development of the UCS repertoire. Secondly, it will enable more efficient processes and reduce latency in publication of approved new repertoire.

During discussion at SG2 meeting #29 in Prague, various participants mentioned concerns regarding the decision-making process and the future role of WG2. This will also be discussed.

1. Introduction

SC2 faces a challenge for long-term, on-going maintenance and development of the UCS—the character repertoire of ISO/IEC 10646: the task of producing the content related to the character repertoire—the code charts and data files—grows steadily larger and more complex, while the capacity of SC2 to perform that task is limited.

Currently, for new characters to be encoded in UCS, they progress through various stages of expert review, NB balloting on drafts of 10646, then publication of a new edition or amendment of 10646. Throughout that process, there is a need to produce charts and data files. The capability to produce charts and data files is critical to this process. As we will explain later, we see a major risk that SC2 will not be able to maintain that capability.

The current process for encoding new characters is the process used for development of new international standards. That process is designed to ensure careful, international review over key design elements of a standard needed to ensure international adoption and interoperability, with inherent latencies built in. As we will describe in more detail below, in the context of 10646, this is essential for some aspects of 10646, such as the basic architecture of the character

encoding, but it can hinder efficient progress in encoding of characters needed by user communities.

Our primary concerns regarding the current SC2 processes for encoding of new characters are those introduced above, and are the reasons why we believe that introduction of a maintenance agency not only would be beneficial for on-going maintenance of the UCS character repertoire but also is essential for its long-term viability.

During discussion of the MA proposal at SC2 #29 in Prague, there were questions and concerns raised as to whether experts from various national bodies would still continue to have input into evaluation of characters proposed for encoding. It seems there may have been a perception that formation of an MA would eliminate that. That is not a necessary outcome from a decision to establish an MA, and we consider expert input important and necessary.

In the following sections, we will elaborate on the above points in more detail.

2. Key motivations for establishing an MA

2.1 Long-term viability of maintaining UCS

The content of ISO/IEC 10646 can be divided roughly into two kinds:

1. the text that establishes the structure of the UCS character encoding, along with its different encoding forms; and
2. the content that documents the coded character repertoire—mainly, code charts and data files.

The first kind of content is largely stable and comprises about 4% of the whole (roughly 100 pages out of over 3000). The second kind of content, that pertaining to the character repertoire, comprises about 96% of the document content, as well as a number of data files. The second set of content will continue to grow in size over time while the first set becomes an ever smaller portion of the whole.

Production of the first set of content from 10646 presents no difficulties: it could be done by any capable editor using any ISO-supported document editing software, including the Online Standards Development (OSD) tool that the ISO/IEC TMB/SMB is encouraging all committees to use. As ISO and IEC continue to explore new tools for editing and delivering standards, it will be easy to adapt this first set of content to those new tools.

The larger volume of content related to the character repertoire is entirely different. Besides being very large, it is very complex, and it becomes increasingly large and complex as the repertoire continues to expand. It is also technically very difficult to produce, requiring specialized tools that are provided by the Unicode Consortium. Moreover, in its current form, as

part of the text of ISO/IEC 10646, it makes this standard completely incompatible with the OSD and other future tools that ISO and IEC envision for their standards.

In addition to requiring specialized tools, production of the code charts and data files in the current process requires special expertise and experience. SC2 has exactly one person with that capability: the current project editor, Michel Suignard, who does this work for both the Unicode Consortium and SC2.

Michel has been in the role of project editor for over 20 years. Over the 20 years, his skills and expertise in production of the UCS content has grown as that content has grown significantly in complexity. While SC2 has named others as contributing editors, it has been the case for many years that production of each edition and amendment of ISO/IEC 10646 has been done single-handedly by him. Having been in this role for many years now, SC2 should anticipate that Michel could soon wish to retire from that role.

A stark reality that SC2 must consider is that, when Michel retires as project editor for 10646, SC2 will not have the capacity, of itself or even with ITTF support, to produce the content for the UCS repertoire. To develop that capacity would require a very large investment in tooling as well as identifying a project editor willing and able to make a long-term (10+ years) commitment to developing the skills and doing the work that is involved, and then training their replacement in the future.

SC2 went through a similar experience recently, when the previous project editor for ISO/IEC 14651 was no longer available: SC2 could not find a new editor to make a long-term commitment to work on that project. The only way for SC2 to create a future for 14651 in which it continues to be relevant as the UCS repertoire grows was to restructure it, separating the evolving data separate from the stable text of the standard, and delegating maintenance of CTT data to Unicode.

If SC2 could not find a long-term project editor for 14651, which is a much smaller and simpler project, it is questionable whether long-term transition of 10646 to a new project editor for 10646 could succeed given the current model of maintenance.

Yet in the meantime, the Unicode Consortium will remain committed to producing The Unicode Standard for the indefinite future, and is equipping itself to do so beyond the point at which Michel might retire. The content related to character repertoire for the UCS is nearly identical to that for the Unicode Standard. The Unicode Consortium will have capacity for development and publication of that content in the future, whereas SC2 will not.

Therefore, to ensure the long-term viability of on-going maintenance of the UCS, a primary goal for a maintenance agency should be to transition responsibility for **development and publication of code charts and data files for the UCS repertoire to Unicode**. By separating the character-repertoire content of 10646 (96%) from the nearly-stable text content (4%), it will

be far more feasible for on-going work on 10646 and the UCS repertoire to be sustained into the future.

2.2 Improving efficiency and reducing latency for publishing new character repertoire

In the current process for UCS, adding new characters to the repertoire uses the ISO/IEC standardization process, which requires a sequence of committee drafts for review and comments by national body experts, followed by one or more enquiry draft ballots for national body approval and comments. Each of these steps has a minimum time period allowed for response (8 weeks for CD consultation, 12 weeks for DIS ballot), with additional time after each for the secretariat to compile results, for disposition of comments to be assessed, and a new draft prepared; or, after final approval, for the new standard to be published by ITTF.

Of course, any process requires time for evaluation, review and revision. Conceivably, different proposals for encoding new characters could progress in parallel, each on its own timetable. The current process does require that new characters be incorporated into a new edition or amendment of 10646, but conceivably, incremental sets of characters could be added in smaller amendments. This was, in fact, done during an earlier period: over four years between 1996 and 1999, twenty-three amendments to ISO/IEC 10646-1:1991 were published! However, ISO Directives have since been changed and now permit at most two amendments before a new edition is required. For that reason, we are compelled to combine more new characters into larger amendments or a new edition in projects that have longer durations.

The timing required using the ISO/IEC standardization process limits agility and forces latency in encoding of new characters. In many cases, this may not be a concern; for example, for historic characters of interest mainly to specialist paleographers, it might be sufficient to know that code points have been chosen in principle so they can progress in digitization projects, and not need to wait for publication of the standard. But in some cases in which international commerce matters, formal approval and publication of the standard is essential. When urgent needs arise, the current process which is not designed for quick processing is unhelpful.

As a recent example, consider China's urgent need for several hundred new CJK unified ideographs, which arose late in 2022. China considered the ISO standardization process too slow to meet their urgent encoding need, and so planned to unilaterally assign characters to reserved code points.[1] Through combined efforts from several CJK experts working apart from IRG, their plan was revised to what later became CJK Extension I, which was added to drafts for

[1] After the first draft of GB 18030-2022 Amendment 1 was published with characters assigned in plane 10 (0xA), the US national body submitted comments recommending that the urgently needed characters be submitted to IRG for processing as a UNC proposal. In their response to the above recommendation that China published with the second draft, they referred to the long approval cycle for international standards ("在国际标准审批周期普遍较长的背景下"). (See L2/23-100.)

Amendment 2 to the 6th edition. But now, over two years since China identified their urgent need for additional CJK characters, Amendment 2 is still not yet published.[2]

In contrast to the current UCS process, the Unicode Consortium has had a yearly cadence for publishing new versions of the Unicode Standard. Unicode's process has still needed to remain in sync with work in SC2 on 10646; that has implied that Unicode cannot add new characters to its standard until there is stable consensus on characters and code point assignments in WG2. But when there is stable consensus, it is able to bring new repertoire into a published industry standard quickly.

For example in the case of CJK Extension I, WG2 had reached consensus on the repertoire and code point assignments by the end of the WG2 #70 meeting in June 2023, which made it possible for Unicode to encode Extension I in Unicode 15.1, which was released in September 2023, less than a year after China's urgent need was made known.

As another example, consider the urgent need to U+32FF SQUARE ERA NAME REIWA. The need for this character arose from the ascension of a new Emperor of Japan on May 1, 2019. That a new character would be urgently needed was known over a year in advance,[3] and a stable code point was determined some time in advance, but the name and appearance of the character was not known until one month in advance, April 1, 2019.[4] Once those details were known, there was an *immediate* need for a published standard to allow international vendors to implement support for the new characters as quickly as possible to align to the May 1 date. In SC2 process, the new character was added to drafts for the 6th edition of 10646, which was not published until December 2020, 19 months after international interchange of the character was required. In contrast, Unicode was able to act with agility, outside of its normal once-per-year cycle, to publish Unicode 12.1 with this one additional character on May 7, 2019.

Therefore, we recommend that maintenance of the UCS character repertoire be separated from the near-stable text of 10646, with updates to the repertoire published by Unicode (as proposed in the previous section) concurrent with releases of the Unicode Standard. That would eliminate problems of latency and bring great agility to maintenance of the UCS repertoire, which it has always needed but lacked.

In an MA structure that formalizes a role for Unicode in the release of UCS updates, the MA terms of reference should provide clear channels for timely feedback from NB or other experts to avoid situations in which useful feedback either comes too late to be taken into account, or can only be taken into account by delaying characters by one full year or more.[5] When

[2] DAM 2 was approved in November 2024, and submitted to ITTF for FDAM processing in January 2025.
[3] See SC2/N4577, submitted in December 2017.
[4] See L2/19-094
[5] For example, CJK Extension H was included in CDAM1, circulated in August 2021. Based on apparent stability, Unicode had included all of Extension H in the "beta" review for Unicode 15.0. But then Japan requested that two characters be removed based on possible unification with other characters. Because of the schedule for IRG meetings, the proposed unifications could not be considered by IRG until after

comments are provided late in the process, it becomes disruptive and problematic for WG2, IRG and Unicode.

2.3 Comparison with other similar standards

Most ISO and IEC standards are limited to content typical of business documents, but some contain supplementary information that doesn't lend itself to text documents. Some standards have large or open-ended sets of entities that evolve over time; for example:

- ISO 639, *Code for individual languages and language groups*
- ISO 1750, *Pesticides and other agrochemicals—Common names*
- ISO 3166 (multiple parts), *Codes for the representation of names of countries and their subdivisions*
- ISO 4217, *Codes for the representation of currencies*
- ISO 5345, *Intelligent transport systems—Identifiers*
- ISO 10957, *Information and documentation—International standard music number*
- ISO 10962, *Securities and related financial instruments—Classification of financial instruments (CFI) code*
- ISO 15924, *Information and documentation—Codes for the representation of names of scripts*

This is just a sample of such ISO standards. All of these separate the information regarding the sets of entities maintained from the text of the respective standard and have that information maintained separately by a registration authority or maintenance agency.[6]

The ISO Directives, Part 2, provides the following guidance on such content:

6.5 Supplementary content

Certain documents contain supplementary material that is best provided in electronic formats, which differs from that of the main content.

EXAMPLE 1 Data sets, code components, test forms.

Unicode 15.0 had been published. At WG2 #69, all including Japan's experts considered the requested change to be quite late and concluded to not accept Japan's request.

Similarly, Telegu and Kannada SHRII characters were included in CDAM2, which was circulated in August 2022, and over more than a year and a half went through three CDAM consultations without comment. Then in CDAM2.4 India requested more time for review, Based on apparent stability, Unicode had planned to include them in Unicode 16.0 but had to remove them late in the development cycle. This delay will mean at least three years pass between when a mature proposal has been vetted by experts and when the characters are encoded.

[6] The distinction between a registration authority and a maintenance agency has to do with the nature of the content to be maintained and its maintenance processes. A list of all ISO maintenance agencies and registration authorities can be found at https://www.iso.org/maintenance_agencies.html.

> This supplementary material can be provided as an attached file or a hyperlink (URN or URL).
>
> Supplementary content shall only be provided in this way if it cannot reasonably be included in the main body of the text.

In the case of ISO/IEC 10646, the code charts and data files that document the coded character are clearly of a distinct nature from the main text. The data files are already kept separate from the text of the standard. The inclusion of code charts in the main text is awkward and inconvenient: it is difficult for users of the standard to consume over 3000 pages of code charts, and preparation of the entire text as if it were a single document is utterly infeasible: the code charts are generated using special-purpose tooling to produce a PDF file, and the PDF files with preceding text, code charts, and following text are spliced together. In comparison with other similar standards and guidance in the ISO Directives, the current structure of ISO/IEC 10646 is an anomaly.

The guidance of Directives, Part 2 leaves open the possibility for code charts to be published separately from the text without requiring a maintenance agency. This is done in the case of ISO 7000, *Graphical symbols for use on equipment—Registered symbols*.[7] That is not a good comparison for UCS, however: those are simply a set of independent PNG files, which ISO Online Browsing Platform (OBP) can easily accommodate. In contrast, UCS is comprised of characters with complex properties and relationships that need to be presented. UCS needs specialized mechanisms that ISO OBP is not designed to support. In a case like this, a separate agency specifically equipped for the type of information involved makes most sense.

3. Involvement of NB experts in review and decision-making processes for new repertoire

During discussion of the MA proposal at SC2 #29 in Prague, there were questions regarding how NB experts can actively participate in the evaluation of characters proposed for encoding. In general, we consider input from international experts valuable and important. For deeper consideration, we should distinguish between CJK unified ideographs and other characters.

In the case of CJK unified ideographs, special expertise is required in the analysis of Han unification. Unicode has depended on the knowledge and experience of experts that participate in IRG. In addition, the processes and the online review tool (ORT) that IRG have in place have proven value. Thus, we believe those experts and processes will remain important for on-going work in encoding CJK unified ideographs, for as long as there are more candidate Han ideographs to be evaluated for encoding.

For non-CJK characters, there is some distribution of expertise. In recent years, the largest portion of expert review and comments has come through the Unicode Technical Committee's

[7] The graphic symbols for ISO 7000 are published on the ISO Online Browsing Platform: https://www.iso.org/obp/ui/#search/grs/.

Script Encoding Working Group (SEW), which meets monthly and in which experts from several countries participate.[8] There are also many occasions in which experts from various national bodies have had input, typically at annual WG2 meetings, particularly when they have been directly involved in developing the encoding proposals. As one participant in the discussions in Prague observed, “We all care about our own characters.” That is to be expected, and we support having NB experts being part of processes for evaluation of encoding proposals.

The main question we see in relation to expert review is what structure and mechanisms should be put in place to facilitate such input. This is a question for AHG1 to consider and for SC2 to determine.

4. Conclusion

We have presented what we believe are strong motivations for establishing a MA for maintenance of the UCS character repertoire, to be maintained separately from the near-stable text of ISO/IEC 10646, and delegating to the Unicode Consortium responsibility for production and publication of code charts.

We also believe that there should be an ongoing role for a broad range of experts to provide input on candidates for encoding, and especially for experts and processes in IRG for the continued coding of CJK unified ideographs. We leave for discussion within AHG1 and SC2 how the terms of reference for an MA should be defined and how those terms affect how experts are to be engaged in processes.

[8] NB experts who would like to join in SEW monthly meetings or contribute to offline discussion of proposals are welcome to participate.