ISO/IEC JTC1 N5___
ISO/IEC JTC1/SC2 N3259
ISO/IEC JTC1/SC2/WG2 N1963

1999-02-22

ISO/IEC JTC1
Information Technology
Technologies d'Information
Информационная технология

Doc Type: **National Body Contribution**
Title: **Response to Japanese National Body recommendation to ISO/IEC JTC1 concerning the activities of JTC1/SC2 (JTC1 N5698)**
Source: **National Standards Authority of Ireland**
Status: **This document is circulated to JTC1 National Bodies for information and review. Copies are circulated to JTC1/SC2 and to JTC1/SC2/WG2**
Action: **ACT**
Date: **1999-02-22**

This document responds to the concerns voiced by the Japanese National Body in JTC1 N5698 with regard to the development of JTC1/SC2's standards, in particular to the development of ISO/IEC 10646, the Universal Character Set. We give the text of JTC1 N5698 and our comments to it below.

> It is Japan's understanding that the work of JTC1/SC2 on the standardization of coded character set is becoming more and more important as the information technology based globalization evolves all over the world. However, Japan has some concern on the activities of SC2 especially from a point of view on the market relevance which was one of the main issues of the JTC1 re-engineering and partly pointed out in the Netherlands contribution JTC1 N5449. Therefore, Japan recommends JTC1 to ask JTC1/SC2 to carefully study the following items and report back to JTC1 possibly by its next to the next plenary meeting.

The "market relevance" of the Universal Character Set is based on its *intrinsic universality,* not on the economic importance of any individual character or set of characters coded in it.

The express scope of this standard is the "representation, transmission, interchange, processing, storage, input and presentation of the written form of the languages of the world as well as additional symbols" (ISO/IEC 10646-1:1993 clause 1). The "market" implied by this scope is no less than the users of *the languages of the world* – it is not intended to be limited to the users of economically dominant languages; nor is it intended to exclude the users of economically less important languages. The Universal Character Set is intended to be the medium for encoding all of the linguistic heritage of humankind. The UCS can be expected to be used for centuries to come. It is difficult to name as easily many other International Standards about which one could say the same. The UCS as a whole has market relevance to the whole world. Whether or not any individual scripts or characters contained within the UCS has particular market relevance is immaterial. The UCS would lose its market relevance – to academic, ecclesiastical, and popular interests – were JTC1/SC2 to endeavour to restrict its work to a narrowly commercial interpretation of "market". To do so would, in our view, be a grave mistake.

> 1. It seems that some of the character sets within the work items of SC2 have quite little market needs compared with other live scripts which are not scheduled for standardization yet. For example, SC2 is working for several dead or almost dead ancient character sets. On the other hand, we know that there are a large number of live scripts or character sets waiting for standardization especially in the Southeast Asia. We also know that ancient scripts are virtually inexhaustible in terms of their repertoires and kinds of characters. Taking these circumstances into consideration, it is apparent that we should be provided with some measure on the market needs for prioritization of the work items for the standardization of character coding.

Irish experts have been actively involved in working to encode the scripts of the world – both "living" and "dead" – together with other experts within JTC1/SC2/WG2, and with experts on the

Unicode Technical Committee. It is true that there are many scripts yet to be encoded, and a group of JTC1/SC2/WG2 and UTC experts has been engaged in mapping out and prioritizing the scripts to be encoded. Careful research, estimation, and planning has been involved here. Two documents, JTC1/SC2/WG2 N1949 (http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1949.pdf) and JTC1/SC2/WG2 N1955 (http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1955.pdf), outline the currently proposed Roadmaps to the BMP of the UCS and to Plane 1 of the UCS respectively.

It is simply not the case that Southeast Asian scripts have been neglected. The active list of such scripts being investigated includes:

Batak (a living script of Indonesia): preliminary discussion document http://www.indigo.ie/egt/standards/iso10646/pdf/batak.pdf
Buginese (a living script of Indonesia): http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1930.pdf
Cham (a living script of Vietnam): preliminary document N1559, discussion document N1960, http://www.indigo.ie/egt/standards/ch/ch.html, and http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1960.pdf respectively
Dehong Dai, Lanna Tai, New Tai Lü, Viêt Thái (four living scripts for Tai languages in China, Myanmar, Thailand, and Vietnam): a preliminary discussion document based on three earlier papers (JTC1/SC2/WG2 N966, N967, N1018) can be found at http://www.indigo.ie/egt/standards/iso10646/pdf/tai-analyses.pdf); discussion of Viêt Thái at http://www.indigo.ie/egt/standards/iso10646/pdf/viet-thai.pdf
Lepcha/Róng (a living script used in Sikkim): preliminary discussion document http://www.indigo.ie/egt/standards/iso10646/pdf/rong.pdf
Ol Cemet', Sorang Sompeng, Varang Kshiti (three living scripts for Munda languages in India): proposals are found at http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1956.pdf, http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1957.pdf, and http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1958.pdf
Pahawh Hmong (a living script of Laos used by emigrés in Australia and N. America): preliminary discussion document http://www.indigo.ie/egt/standards/iso10646/pdf/hmong.pdf
Siddham (an ecclesiastical script used especially in Japan): preliminary discussion document http://www.indigo.ie/egt/standards/iso10646/pdf/siddham.pdf
Tagalog, Buhid, Hanunóo, Tagbanwa (one dead and three living scripts of the Philippines): http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1933.pdf (The proposals for these scripts have been endorsed by the UTC and await approval by JTC1/SC2/WG2.)

In addition, the Roadmaps to the BMP and to Plane 1 list for future consideration the following living Southeast Asian scripts: Javanese, Kayah Li, Kirat/Limbu, Lisu, and Manipuri/Meitei, as well as the following dead ones: Balinese, Chalukya, Jurchin, Kaithi, Khamti, Kharoshthi, Kitan, Pyu, Rejang, Satavahana, and Tungut.

Of the many historical scripts proposed for addition to the UCS, two things may be said. First, several of them have considerable bodies of literature and are of great relevance to the linguistic and cultural heritage of the world. Egyptian hieroglyphs, Sumero-Akkadian cuneiform, and Mayan hieroglyphs are perhaps most relevant to European and American interests; Plane 2 has been dedicated to CJK extensions to meet Asian needs. Second, active work on some of these scripts has been possible because they are small, simple, well-attested, and straightforward to encode, and because expert resources have been available to work on them (Ogham, Runic (already encoded); Byzantine Musical Symbols, Etruscan, Glagolitic, Gothic, Western Musical Symbols (endorsed for encoding in Plane 1); Avestan, Old Hungarian, Old Permic, Phoenician (under discussion). If Southeast Asian scripts have received less attention from active experts, it is generally because access to expertise and source materials has not been available. We would encourage the Japanese National Body and other National Bodies to help overcome this shortfall by providing access to such expertise and source materials if possible.

> 2. If it is true that the virtually inexhaustible sets of characters must be standardized, it will be quite difficult, in terms of expertise and resources, for JTC1 to develop and maintain all these standards.

In the first place, the sets of characters, whether representative of living scripts or dead ones, is not inexhaustible. In the second, JTC1 does not itself do the work of encoding. Neither does JTC1/SC2. It is experts in JTC1/SC2/WG2 who do this work, in close cooperation with experts in the UTC.

> In reality, JTC1 will have to focus on the fundamental and kernel part of the standardization of character coding asking some reliable outside organizations with appropriate expertise to take care of the rest of the standardization work.

JTC1/SC2/WG2 and the Unicode Technical Committee engage in such consultation as a matter of course. For example, representatives of Sri Lankan and Myanmar standardization institutes were consulted as the encoding standards for the Sinhala and Myanmar scripts (presently closing their ballots) were developed. Likewise, the Center for Computer-Aided Egyptology in Utrecht has been consulted in the development of the proposal to encode Egyptian hieroglyphs in Plane 1 of the UCS (see JTC1/SC2/WG2 N1955, `http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1955.pdf`).

> To make these distributed development and maintenance possible, we must also be provided with a clear and detailed standardized framework for character coding with recognition and registration mechanisms to attain and maintain interoperability.

Just such a framework is employed by JTC1/SC2/WG2 and the UTC, in the standing document on Principles and Procedures for Character Encoding (N1502R, N1876, and related documents), developed jointly by the two committees. If there are shortcomings in this document, the Japanese National Body may be able to provide suggestions on how it might be improved. We note that a proposal to add more precision to the “script category” described in this document has been offered by the Netherlands in Attachment C of JTC1 N5449, and we suggest that this proposal be recommended to the Ad-hoc maintaining the Principles and Procedures document for possible inclusion.

This comment from the Japanese National Body does cause us some concern, however, if the implication is that responsibility for encoding and maintaining any part of the UCS be delegated to bodies other than JTC1/SC2/WG2 and the UTC. Doing so could jeopardize the integrity and excellence of the UCS. External bodies and experts are always to be consulted to insure that their needs are met – but the stringent architectural and structural concerns must continue to be controlled by WG2 and the UTC.

The Universal Character Set is *not* a registry. It is a coded character set. The ISO 2375 registry exists to provide a mechanism for encoding character sets outside the UCS, should any organization prefer to do so in lieu of including their characters in the UCS.

> 3. The current and future basis for the standardization that satisfies various needs including above requirements will be ISO/IEC 10646. However, the basic concept of IS 10646 seems to be changing and ambiguous.

We do not believe that the “basic content” of the UCS is particularly ambiguous. It is clearly stated in the scope of ISO/IEC 10646. Some members of JTC1/SC2/WG2 are concerned that the current JTC1 procedures require explicitly 5 member bodies to agree, formally to “participate in the development” of a given script, at the time when a subdivision is proposed to allow such a script to progress to PDAM. In practice, however, a script is proposed for PDAM *only* after a strong consensus among experts involved has been achieved – often, as in the case of Khmer, Myanmar, Sinhala, Thaana, and the Philippine scripts, with prior decision by the UTC to endorse the proposal. What this means is that the danger exists that a perfectly acceptable and encodable script could fail to be added to the UCS for purely formal reasons. We do not (necessarily) suggest changing the procedures, because, as the Netherlands pointed out in Attachment C of N5499, “dealings like ‘if you participate in my project on a script you do not understand, I’ll do in yours which I do not understand either’” have in fact been employed on occasion by JTC1/SC2/WG2. This does *not* mean that WG2 endorses a “loosening of criteria for participation”; it merely means that WG2 recognizes that, in the special case of the UCS, the formal requirements for a project subdivision do not take into account the fact that, at the time WG2 has enough confidence in a proposal to recommend

project subdivision and PDAM ballot, the development work has *already been completed.*

For example, introduction of control tag structure is being considered recently, and the plane assignment is given for limited number of planes in rather ad hoc manner. We must be provided with a clear guiding principles for the future extension of IS 10646.

The definition of the use of Planes external to the BMP (10646-1) is a matter for the scope of ISO/IEC 10646-2, which is currently under development. National Bodies concerned with the precision of the definition of these planes can make recommendations and propose text to aid the development of that part of the UCS. The concern of the Japanese National Body with regard to this issue is a valid one, and attention should be given to the question by WG2 and the UTC.

It is the view of the Irish National Body that JTC1 should commend JTC1/SC2 and JTC1/SC2/WG2 for the excellent work they have done and continue to do, and for the close and effective cooperation they enjoy with their industrial counterpart, the Unicode Technical Committee, which maintains the Unicode Standard.