L2/03-322

ISO/IEC JTC 1/SC 2/WG 2 *N2635*
Date: 2003-09-30

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
ISO/IEC JTC 1/SC 2/WG 2

Title: Comments on N2621
Source: Robert R. Chilton, Asian Classics Input Project
Status: Individual contribution
Action: For consideration by WG2

======= POINTS TO MAKE WITH REGARD TO N2621

1. The proposed set of precomposed BrdaRten characters is both insufficient and unnecessary. This character set does not encode a script but rather a limited set of glyphs; whereas comprehensive encoding for Tibetan-script orthography is already accomplished via the 193 characters of the 0Fnn block.

Result: added complexity without any real benefit; users who need comprehensive encoding of the Tibetan script will not have their needs met via the BrdaRten character set.

2. Consistent (and culturally expected) data processing of Unicode Tibetan materials (e.g., in searching and sorting) will require that any precomposed BrdaRten characters be decomposed (normalized) into equivalent 0Fnn character strings.

Result: increased difficulty of implementation and substantially increased processing load at runtime.

3. This proposal for precomposed Tibetan ligatures, like those previously submitted as N964 (1994) and N2558 (2002) must be intended to address uses of Tibetan script in systems that do not support smart fonts.

Result: assumes that support for complex scripts (smart fonts) is unnecessary, which seems to run counter to the general UCS model; worldwide users who need to work with complex scripts (and therefore use smart fonts) will have no reason to use these precomposed BrdaRten characters.

4. Because of the problems and limitations of the BrdaRten character set described in items #1 to #3 above, we can fairly assume that virtually all users of Tibetan script outside of the PRC will employ smart fonts for Tibetan (using the characters encoded in the 0Fnn block) and will completely avoid using the BrdaRten character set.

Result: encoding the precomposed BrdaRten characters would create a second encoding model for Tibetan-script materials; implementations that claim to support Tibetan would need to support both encoding models and would therefore be significantly more complex than if just a single encoding model for Tibetan (i.e., relying on the 193 characters of the 0Fnn block) is used universally.

======= RESPONSE TO N2558

ABSTRACT

The main objection to n2558 is that it is simply unnecessary; the existing ISO-10646 Tibetan character set is not only sufficient but enables a far greater range of Tibetan-script orthography than the character set proposed in n2558.

Moreover, for the authors of n2558 to argue that a non-combining model of Tibetan is necessary for compatibility with "traditional education, publication and electronic desktop publishing systems" to is to entirely discount the use of other complex scripts -- such as the Indic scripts which employ a combining model -- in such "systems". Clearly, the direction of such a rationale runs entirely opposite to the basic principles of ISO-10646.

Acceptance of this proposal would introduce an alternate encoding of Tibetan-script data. This would increase processing load and complexity due to the need of continually normalizing between these two differently encoded Tibetan-script data sets. Thus, the negative impact of this proposal within the overall context of the ISO-10646 character set would far outweigh any supposed benefits.

DISCUSSION

It seems that this proposal is motivated largely by typographical considerations without proper concern for broader character data processing needs. Although this character set might be fine for computer-based typesetting of the Tibetan materials now being printed in the Peoples' Republic of China, it is inadequate as a basis for interchange and processing of Tibetan-script data.

Most notably this proposal represents the repertoire of a particular usage (Tibetan as currently used in the Peoples' Republic of China) rather than a script. There are many examples of Tibetan-script words in classical Tibetan works, as well as in Dzongkha and

other Tibetan-script languages of South Asia, that cannot be represented by this character set.

Secondly, if the goal of this proposal is to facilitate processing of Tibetan-script data for purposes other than document publishing then it would have been more effective to provide characters for every Tibetan initial form (including prefix letters) rather than simply for typographical ligatures. The proposal as now written will result in unnecessary complexities in producing a culturally expected collation of data encoded using mixed basic Tibetan and BrdaRten characters.

More specifically, the proposal contains some errors of fact:

1. The claim that "[the current Tibetan-script] encoding scheme is not compatible with traditional education, publication and electronic desktop publishing systems" is simply not true. Any system that is able to render other complex languages, notably the various Indic and Indic-derived scripts of South and Southeast Asia, should be able to accommodate Tibetan-script materials encoded using the current Tibetan block. (It is no coincidence that the Tibetan script, which is itself derived from ancient Indian script, should share many structural and functional characteristics with modern Indic scripts.)

It is understandable that the authors of n2558 would like to regard Tibetan as "a horizontal stream of basic Tibetan characters and BrdaRten characters without vertical combining" since this facilitates the usage of two languages, namely Tibetan and Chinese, together in bi-lingual documents. However, this mode of thought runs counter to the very principle of ISO-10646 which is to enable *any* number of languages to be used together, seamlessly, in documents and other computer applications. Would the authors of n2558 also like to propose a set of precomposed characters for each of the Indic scripts so that they likewise can be "regarded as a horizontal stream of [basic Indic] characters and [precomposed Indic] characters without vertical combining"? Or have they resigned themselves to never mixing Chinese and Indic script within a document? On the other hand, once there is a system that can render Chinese together with Hindi or Tamil, rendering of Chinese together with Tibetan (as currently encoded) is not technically difficult.

In point of fact, the cited "problems with Tibetan information interchange and processing" are no more difficult to solve than those for other complex scripts -- these having already been solved for a substantial number of complex scripts. The current lack of widespread support for ISO-10464-standard Tibetan simply reflects the fact that there are fewer commercial and governmental resources being allocated to the development of ISO-10464-standard Tibetan as compared to other Indic and Indic-derived complex scripts.

2. The claim that "Up to now, there is no report showing any system platform has implemented Tibetan processing system using dynamic combining method" is also untrue. [Refer to the recent communication from Steve Hartwell re: N2621 for details.]

3. The statement that "Since 1990s, from DOS to Windows, both domestic and overseas applications have been using Tibetan BrdaRten character set at implementation level 1. For example, the Founder desktop publishing system for Tibetan is based on BrdaRten characters which has become the de-facto industry standard for Tibetan information interchange and processing in China and even outside of China" is exaggerated. Tibetan-script computer systems have been in use in North America, Europe, South Asia and East Asia/Pacific Rim as early as 1983 but it is completely false to say that the character repertoire of n2558 has become "the de-facto industry standard for Tibetan information interchange and processing" in any place outside of China. As noted above, the character set of n2558 does not even fully support usages of Tibetan script in regions outside of China. (The notation of "Worldwide" in question 5 of the Part C.: Technical-Justification in the Proposal Summary Form is thus highly misleading.)

4. The n2558 document asserts that "Once the Tibetan BrdaRten characters are encoded in BMP, many current systems supporting ISO/IEC10646 will enable Tibetan processing without major modification. Therefore, the international standard Tibetan BrdaRten characters will speed up the standardization and digitalization of Tibetan information, keep the consistency of implementation level of Tibetan and other scripts, develop the Tibetan culture and make the Tibetan culture resources shared by the world." There are a number of counter-arguments to these assertions:

First, due to the limitations of the n2558 character set for representing classical Tibetan, Dzongkha, and other Tibetan-script materials it is not reasonable to expect worldwide adoption of this character set. Since the dynamic-combining model will continue to be used in South Asia (where complex-script systems are the norm), in academic institutions (where research in classical Tibetan is conducted) and elsewhere, there will always be a need to normalize Tibetan-script data interchanged between regions that use these two differing encoding models for encoding Tibetan-script data. Thus, the acceptance of this character set into the ISO-10646 standard will actually be an *obstacle* to "standardization and digitization of Tibetan information."

Second, the reference to "consistency of implementation level of Tibetan and other scripts" would seem to presume that the "other scripts" in question are not complex scripts. This statement is simply not relevant when we consider the requirements of -- and the already implemented multilingual systems for the handling of -- Indic and Indic-derived complex scripts.

5. Any claims of a pre-existing "de-facto industry standard" for Tibetan even in China seem to be contradicted by the statement in the Conclusion, that "After serious discussion and analysis by Tibetan linguists, encoding experts and software developers in China, all are in favor to establish a national and international standard Tibetan BrdaRten character set to meet the requirement of Tibetan information processing." This seems to indicate that a national standard for Tibetan is yet to be established, even in China.

In summary assessment, had this proposal been comprehensive enough to satisfy the needs of *all* users of the Tibetan-script languages and materials, had it taken into

consideration character data processing needs of Tibetan beyond computerized typesetting (such as collation), and had it been presented ten years ago, then it might have well been worthy of serious consideration. As it now stands, this proposal offers too little too late and, moreover, would simply add further confusion and obstacles to the standardization of Tibetan-script data processing and interchange. Furthermore, even had this proposal been presented for consideration ten years ago, the fact that complex-script (dynamic combination) rendering is needed for Indic scripts would even then have been a strong argument in favor of the current ISO-10646 encoding model for Tibetan script and against an encoding model of the type proposed in n2558.

Respectfully,

Robert Chilton

As for my qualifications:

1. I have overseen technical issues at the Asian Classics Input Project (ACIP) since 1993, especially with regard to conversion of the ACIP text database, currently more than 165,000 pages of Tibetan e-text and catalog data, to Unicode.

2. As early as 1990 I co-authored a computer program for sorting Tibetan. I am very familiar with the issues involved in searching and sorting Unicode Tibetan, having recently been invited to present a paper at the Tenth Seminar of the International Association for Tibetan Studies (Oxford University, September 2003) entitled "Sorting Unicode Tibetan using a Multi-Weight Collation Algorithm".