ISO/IEC JTC1/SC2/WG2 N3079

L2/06-118

2006-04-08

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type:	**Working Group Document**
Title:	**Response to UTC contribution N3069, "Concerns Regarding WG2 N3043R, Myanmar Additions to 10646"**
Source:	**Irish and UK National Bodies**
Status:	**National Body contribution**
Action:	**For consideration by JTC1/SC2/WG2**
Date:	**2006-04-08**

The Irish and UK National Bodies have become aware of some concerns expressed by the Unicode Consortium (N3069) regarding the proposal to add seven Myanmar characters to the UCS (N3043R). The present document is part of an ongoing response which has been made between the UTC and the proposers, in which the proposers have consistently provided information and response to issues and questions raised by the UTC, implementors, and National Bodies.

The Irish and UK National Bodies appreciate the general *caveat* offered by the UTC in N3069 regarding stability. Disunifications of the sort proposed in N3043R are not insignificant. But neither have we and the Myanmar organizations who co-authored N3043R made either the proposal or the request for fast-tracking the seven characters precipitously or without good technical cause.

History of Proposal Development

WG2 N3043R is the culmination of an ongoing discussion about how the Myanmar script should be encoded—a discussion dating back to 1998 (N1883R2). More recently the discussion was moved forward with the presentation of document L2/05-184 (by Martin Hosken) to the UTC for consideration. At their August 2005 meeting the UTC minuted: [104-A54]

> **Action Item** for Eric Muller, Rick McGowan: Correlate and analyze the requests in all current Myanmar documents. Come up with a set of questions and concerns, and then invite the Myanmar national body experts to participate in an e-mail discussion. [L2/04-273, L2/04-198, L2/05-178, L2/05-184].

A discussion e-mail list was created but no questions were forthcoming. Instead a meeting was arranged in Yangon, Myanmar 13-15 February 2006 in which the issues were raised. This was an historic event to which key Myanmar implementors known to the Myanmar Computer Federation were invited. To bring such a meeting together with such a united result and an agreement by all of the Myanmar script implementers to hold off releasing implementations which were ready to be released until after the next WG2 meeting, is nothing short of a miracle. The support within Myanmar for this proposal is unanimous. This should not be underestimated.

Since the meeting occurred the week after the last UTC meeting, it was not feasible to submit the proposal to the UTC prior to the WG2 meeting, though the document was circulated to both committees as N3043R (L2/06-077R). The UTC has been kept well-informed and an intense discussion ensued following the publication of the proposal. From this discussion documents N3029, L2/06-085 and N3061 (L2/06-093) were written. To this date no response has been received to N3061, which was specifically written to allay fears, to answer questions, and to resolve issues arising in the discussion. This is evident by the fact that N3069 does not refer to either N3029 or N3061.

Responses
In this section we review and respond to each of the points raised in N3069.

Character disunifications need extra careful review: The discussion on the Unicode Technical Committee's discussion list has been fast and furious with a number of documents appearing, but at no point has any alternative solution to the problem been presented that would do all of the following:

- Simplify the encoding rather than make it more complex
- Not require transcoding (we will return to this)
- Actually address the issue of minority language extensions

Various approaches are discussed in N3061 and dealt with there.

The need for transcoding: What is amazing is that even though the Myanmar script has been encoded in the UCS since Amendment 26 which was published in 1999, there is so little data using it. In fact, nearly all the data known to be in UCS encoding was represented by its creators at the Yangon meeting and all expressed their willingness to transcode. The amount of UCS-encoded data pales into insignificance against the mountain of legacy encoded (8-bit) data that needs to be transcoded. The reason why those at the meeting were willing to transcode was that for many the original text was converted from legacy encoded data in the first place. So transcoding would simply involve changing their existing mapping and reconverting. Another issue is the fact that there is a sizeable body of non-conformant UCS data in existence. A Google search yielded some 30,000 web pages that use a UCS-based font which utilizes unassigned character positions. All of that data would have to be transcoded anyway just to make it compliant—so the burden of transcoding is not something that can be avoided. (The author of the non-standard font is a Myanmar implementor who has agreed to provide free tools to help people transcode their data to the proposed encoding.)

This does not cover all the data and there are some users who are creating original Unicode data. The only known commercial system is from XenoType Technologies and its proprietor, Mr Kaʻōnohi Kai, has kindly offered to help transcode the data of his customers and to transition to using the resulting encoding. It should be noted that XenoType Technologies have some four dozen (that is, 48) customers whose data needs to be transcoded.

What is even more compelling than the paucity of existing UCS-encoded Myanmar data is the fact that all existing users with UCS Myanmar data will need to transcode whatever is decided. N3061 shows that the current Unicode specification for the Myanmar script is under-specified, given that there are different ways of handling *kinzi,* and no consistency in this area among any of the implementors. Resolving that consistency will require transcoding in any case. So regardless of whether N3043R is accepted or not, data is going to have to be transcoded. There is *no solution* that removes the need to transcode.

Major Software Vendors: Given that following the Yangon meeting only *one* company (XenoType) has admitted to having a fully Unicode conformant commercial solution available and that all other implementors have either not released their work or have specifically held back their release to await the outcome of the WG2 meeting, we have to ask who these major software vendors are to which N3069 refers. Participants at the Yangon meeting included people representing at least three different and *conformant* implementations. The consequences on implementation raised by N3043R were well understood and since N3043R constitutes a simplification (and opens the way for support of minority languages), those consequences were welcomed. In addition it must be said that N3043R was not written to make implementors' lives easier since three conformant solutions of the existing standard are waiting in the wings.

Only one implementation exists: While N3043R inadvertently gave the impression that SIL is the only one with an implementation, the proposal makes the issue clear in the rationale for fast-tracking:

> While all current implementations are at research level, some are ready for production and delaying a change would result in considerable text to transcode.

This acknowledges that there are implementations in existence. The XenoType Technologies product was not known to the proposers at that time and immediately it was heard about, the proposers made contact with the implementor and brought him into the discussions.

In addition, the issue of whether "only one implementation exists" or not completely misses the main reason for N3043R being written. It is possible that the Myanmar encoding could be fixed so that it is sufficient for representing the Burmese language without adding these seven extra letters. N3043R never denies this and it has never been denied by any of the proposers. What N3043R does state is that the need for the changes proposed comes from the need to support *minority languages* (specifically S'gaw Karen) and as a side issue the changes also resolve the existing ambiguities in the Myanmar script encoding for Burmese which are currently there due to its under-specification. This has been repeated time and again and has been often forgotten or ignored in the discussions.

N3069 Conclusion: N3069 suggests a few alternative mechanisms that might be tried for resolving the problems stated in N3043R and N3061. Language sensitive font technology is addressed in N3061. Variant selectors could be considered an option but this fails to make things simpler and still does not remove the need for transcoding. It is also a very similar solution to that addressed regarding ZWJ & ZWNJ Profusion in N3061.

The national bodies of Ireland and the UK have requested the fast-tracking of the disunified characters because of industry pressures within the Union of Myanmar. There is an urgent need for a solution that will allow for the plain-text representation of the languages of the Union, and the software industry in Myanmar is ready and willing to implement the solution proposed in N3043R. We believe that WG2 should accept this proposal in order to serve the main users of the script, who co-authored the proposal, and who wish to see it implemented in the standard at the earliest opportunity.