L2/03-368

JTC1/SC2/WG2 N2668

Title:	Comments on N2661, Clarification and Explanation on Tibetan BrdaRten Proposal
Doc. Type:	Expert contribution
Source:	UTC/L2
Date:	October 20, 2003
Action:	For consideration by JTC1/SC2/WG2, UTC
References:	WG2/N2621, N2558 (= L2/02-455), N2624 (= L2/03-315), N2625 (= L2/03-316), L2/03-002r, N2638 (=L2/03-328), N2661
Distribution:	WG2 members, UTC members

Background

In document N2558, the Chinese national body proposed the addition of 956 characters for Tibetan pre-composed "stacks". This proposal has been re-submitted in document N2621 with only very minor changes. After several contributions submitted in response to N2621 raising concerns with the proposal, the Chinese national body submitted a response intended to further justify the proposal in N2621.

The proposal in N2621—specifically, the use of precomposed Tibetan character—is referred to herein as the *BrdaRten* proposal.

Comments on responses to objections provided in N2661

The response to critiques in N2661 is divided into three parts. The first asks whether the BrdaRten proposal is necessary, and answers that question in the affirmative on the basis of history of usage beginning with typewriters, and including software implementations from the early 1990s using proprietary encoding systems. What this section does not address, however, is that decisions regarding encoding models in the UCS cannot be based solely on existence of legacy implementations or the number of users of those implementations. Rather, such decisions must be based on *feasibility and cost* of implementation of the UCS, including interoperability with existing implementations, which in turn must include both legacy implementations and implementations of what is already in the UCS. We have no disagreement with the statement in N2661, "This huge amount of e-data with mission-critical users is another fact which can not be ignored." Such statements do not, however, constitute sufficient grounds for adopting the BrdaRten proposal.

The second section in N2661 asks whether the BrdaRten proposal is sufficient. It answers that question by saying that it is sufficient for Modern-Tibetan-language texts, but not for classical texts; other languages using Tibetan script are not discussed. While acknowledging limitations for classical texts, N2661 does not discuss the implications of mixing the precomposed BrdaRten model with the existing dynamic-composition model in implementations that need to support classical texts. This is a serious oversight in the argumentation of N2661, since the problems of having to mix encoding models would be quite significant, as was already stated in the contributions to which N2661 was supposed to be responding. Thus, these objections to the BrdaRten proposal cannot be considered to have been addressed in N2661.

The third section of N2661 attempts to evaluate the pros and cons of the BrdaRten proposal in comparison with the existing dynamic-composition model of the UCS. The primary argument against the dynamic-composition model in the previous Chinese contribution, N2621, was that the difficulty in creating rendering implementations using that model was considered too great. N2661 extends the case for the BrdaRten proposal by means of presumed implementation and migration costs in several areas for the dynamic-composition model. Each of the ten arguments of this nature will be considered in turn:

1. **Claim:** Text using the dynamic-composition model "is not readable by ordinary people" when there is no complex-rendering support or "smart" font (e.g. OpenType) available, and it is assumed that this will occur on occasion.

   Evaluation: It is true that text in such scenarios would be difficult to read. This problem is certainly one of degrees: obviously, text using the BrdaRten model would be impossible to read in the absence of any Tibetan font on the user's system. The question what minimum amount of support should be required in an end-user application to ensure legibility of text. N2661 assumes that a font should be included, but that complex-rendering support should not.

   It is difficult to see how this argument could be used to justify the BrdaRten proposal. Clearly, in general the evaluation metric assumed by N2661 is not applied for numerous other scripts in the UCS: for every other Indic script, as well as Arabic, Hebrew and others, both a font and a complex-script rendering engine are assumed to be necessary to provide legible display of text. (The same is also true, by the way, for Classical Tibetan texts, for which the BrdaRten proposal would be inadequate.)

   While the authors state that "China does not intend to challenge WG2 general principle on pre-composed character encoding," and that they wish to raise questions only in relation to the Tibetan script, it is clear that changing the existing encoding model for Tibetan in the UCS could be seen as an invitation to other parties to request that the encoding models for other scripts be revised to use solely precomposed or presentation forms. This would be catastrophic for the UCS, and would lead to complete failure in the goal of maintaining interoperability.

2. **Claim:** migration of the large volume of existing legacy data, which uses a precomposed encoding model, which be much more costly for the dynamic-composition model in the UCS because (a) one-to-many mappings are required, and (b) variable-length strings would result.

   Evaluation: No explanation is given as to why one-to-many mappings are thought to be more costly to implement; whether considering cost of development, cost of system resources, cost of processing time, the differences are surely insignificant.

As for the objection that variable-length strings would result, this can be countered from two perspectives. First, the majority of implementations already have to deal with multilingual data having variable length due to the variable-length encoding inherent to the UTF-8 and UTF-16 encoding forms, the result of which is that internationalized applications cannot assume a fixed data width for multilingual strings. Moreover, internationalized applications already have to deal with the fact that graphemes in many languages may vary in the number of characters required to support them. This concern can only be considered applicable in specific scenarios involving legacy applications intended to support only Tibetan text that are being revised to support UCS encoding.

Secondly, it has already been observed that the BrdaRten implementations would still need to support the dynamic-composition model for existing UCS data as well as for Classical-Tibetan data, and hence this problem cannot be avoided in the general case. Thus, scenarios in which this concern is applicable must be further restricted to exclude support for any language other than Modern Tibetan or existing text encoded using the existing UCS encoding. While there may be some who need to implement for such constrained scenarios, the savings for such situations do not compare with much greater costs to others that result from the existence of data in two different representations and the related normalization issues.

3. **Claim:** The dynamic-composition model makes editing operations—cursor movement, deletion, etc.—more difficult.

   Evaluation: Such issues are no different for Tibetan than for other scripts involving combining marks, notably other Indic scripts, and are already addressed in various existing implementations. For instance, the Uniscribe rendering engine recognizes Tibetan stacks as clusters and assists the application in selection, hit-testing (mouse position to string position mapping) and similar processing. Providing a user experience in which stacks are treated as indivisible units, as would result from the BrdaRten proposal, requires only minor additional effort by software developers using the dynamic-composition model.

4. **Claim:** Data using the dynamic-composition model requires greater storage space than data using the BrdaRten model.

   Evaluation: The cost of storage has been decreasing much faster in recent years than the rate at which users generate new text data.

5. **Claim:** The BrdaRten proposal allows simple keyboard input methods, while the dynamic-composition model requires complex input methods using multiple alternate keyboard "pages".

   Evaluation: Any stack that can be entered as a single keystroke in the BrdaRten model can similarly be entered as a single keystroke in the dynamic-encoding model. The only difference is in the number of characters generated by a single keystroke.

   For instance, a layout intended for use with the BrdaRten model could easily be implemented for use with the dynamic-composition model using input-method development tools such as Tavultesoft Keyman or the Microsoft Keyboard Layout Creator. Rather than adding several new characters to the UCS and introducing a competing encoding model, what may be more useful is to implement an input method for such a layout for use with the existing UCS characters.

6. **Claim:** OCR implementation is more difficult for the dynamic-composition model since it requires recognition of individual components within a stack while the BrdaRten proposal does not.

   Evaluation: As was the case in relation to input methods, an implementation that works in terms of the units provided in the BrdaRten proposal can easily be implemented to give the same behaviours using the dynamic-composition encoding model.

7. **Claim:** In post-OCR proofing, a one-to-one correspondence between image and encoded character, as provided by the BrdaRten proposal, is advantageous.

   Evaluation: No explanation is given as to how a one-to-one mapping is supposed to be advantageous over a one-to-many mapping. This argument is, therefore, difficult to evaluate.

8. **Claim:** Sorting using the dynamic-composition model appears to be simpler than using the BrdaRten model.

   Evaluation: Tibetan sorting can be implemented using either model, but will indeed be simpler using the dynamic-composition model since individual components within a stack must be compared. A sorting implementation for use with the BrdaRten model would likely use normalization to convert such data to data that uses the dynamic-composition model. (Such normalization would also be necessary in order to compare data encoded using the BrdaRten model with data encoded using the existing dynamic-composition model.)

9. **Claim:** For searching, there is no difference between the dynamic-composition and BrdaRten models. The example given in N2638 of searching for KA is irrelevant since “such searching is semantically meaningless”.

   Evaluation: The statement that searching for KA is semantically meaningless is unclear. A text element KA may not carry semantic meaning, but that does not mean that users may not find use in being able to search for such a text element. For instance, in preparing a dictionary, a user may wish to filter data to identify all wordforms in a sorted list that would appear under a heading for KA.

10. **Claim:** The dynamic-composition model requires a complex-script rendering engine and advanced fonts, while the BrdaRten model does not.

    Evaluation: This claim is true, but complex-script rendering support and advanced fonts are already required for data using the existing encoding model, and as shown in N2638, the problem of developing such capabilities has already been solved in existing software implementations on various platforms.

N2661 has attempted to show that several different text processes involved in providing complete support for Tibetan-script text are much more complex or costly using the dynamic-composition model than with the BrdaRten model. In response, it is argued here that, in most cases, there is no significant difference, and that in some cases the dynamic-composition model is favoured. The primary argument for the BrdaRten model remains the need for complex-script rendering support with the dynamic-composition model in order to provide legible text. N2661 acknowledges the working implementation of such complex-rendering support for Tibetan demonstrated in N2638, but dismisses it on the grounds that this particular implementation is in development rather than in a shipping product. It does not, however, address the fact that this is not the only working implementation, or that the platform vendor whose implementation was demonstrated has successfully delivered working implementations in shipping products for several complex scripts.

N2661 has not addressed the serious issues that were raised in N2624, N2625, L2/03-002r and N2638 in relation to the BrdaRten proposal leading to alternate representations for Tibetan data. Moreover, it does not address the serious jeopardy in which the UCS would be placed as a result of introducing an alternate precomposed/presentation-form encoding for Tibetan and providing a precedent for other scripts to follow suit.

Conclusion

N2661 has attempted to respond to concerns and to provide a stronger case in support of the BrdaRten proposal, but in fact has not provided significantly stronger arguments than were presented in N2621. Serious problems related to the creation of alternate representations for Tibetan remain, and have not been addressed by the proposers. Accordingly, the recommendation that the proposal presented in N2621 be rejected is upheld.