Response of IP to Bengali LGR Draft of August 9th 2018

DATE: 2018-08-14

# Overview

The IP received and reviewed the following files

Bangla\_LGR\_20180809.doc

Bangla\_LGR\_20180809.pdf

Bengali\_LGR\_20180809-IP2.doc

proposal-lgr-bangla-20180809.xml

test-label-bangla-20180809.txt

and has the following comments. The IP would like to see these (and outstanding earlier comments) resolved before making a final recommendation on whether the LGR is ready for public comment.

# Comment #1

There were a large number of detailed editorial suggestions on the previous draft, both for the DOCx file and for the XML file. A quick check by the IP revealed that these appear to not have been addressed by the GP, in contrast to most (but not all) of the itemized comments which seem to have resulted in changes in the proposal.

The IP has attached a list of previous itemized comments that are still open for the GP to review and respond or make fixes to the LGR where appropriate. Particular note should be taken of issues around the WLE rules, which, if not resolved, may prevent the IP from accepting this proposal.

In addition, this review package includes two marked-up proposal documents; one for the 0626 version and one for the latest one. Both contain editorial corrections and suggestions that the GP is requested to review. Finally, there’s an older XML file containing some suggestions not yet acted on.

# Comment # 2

Clearly there is an unresolved issue on the correct name for this script. It is used to write many languages (the two greatest being Bangla and Assamese) and has a history distinct from that of any one language. However, it was officially named “Bengali” (by ISO Standard 15924, which is the name also used in the Unicode Standard). There were, however, recent discussions by ISO (28 June 2018) about possibly changing the header of the code chart shown in Unicode it to Bengali-Assamese.

Given the multilingual use of the script, it is desirable that “Bangla” be kept for reference to the language of that name, whereas Bengali-Assamese (shortened to Bengali) refers to the modern script.

Note that for consistency with other scripts in the Root Zone LGR and because of **automatic text generated from the script ID (und-Beng) used in the final integrated LGR**, the integrated LGR will need to refer to this script as “Bengali” – as that happens to be the name mapped to “und-Beng” in the Unicode Database.

## Character names

Finally, in making the edits to the latest proposal, the GP changed many names that are externally defined. For example, all formal character names, being immutable in Unicode, retain the spelling BENGALI, and should be referred that way not just in tables but also in the running text of the LGR.

Please see the attached marked-up document for suggestions where “Bengali” should be retained.

This issue also affects the XML. Please review the Description and title and make sure the script is referred to as Bengali and any character names match those used in the Unicode Tables.

Note that in “comment” attributes for characters, the convention is to designated aliases using “=” followed by the other name in lowercase (except for language names), for example:

comment=”= Assamese letter wa”

The IP encourages the GP to use this convention to help document differences in naming between the different communities.

# Comment #3

About Section 3 in the proposal document: This description is about the Bangla language rather than the Bengali-Assamese script. Despite the title “& Principal Languages Using It” only the history of one language, Bangla is traced.

It should be re-written from the point of view of the origins of the current Bengali-Assamese script, mentioning Bangla, Assamese, Manipuri languages etc as their areas of use become relevant.

Therefore, the first paragraph of Section 3.1 and most of 3.2 (especially towards the end) contain the primary content that is needed.

Note that the Procedure requires that the Root Zone LGR be free of bias towards or against any particular language. This is especially important for scripts that are used with multiple languages.

Therefore, the IP would consider it appropriate to start out the description emphasizing the use of the script for multiple modern languages.

In this context, the paucity of descriptions of features and usages relevant to Assamese are troubling as well.

# Comment # 4

In 3.3.8 Zero Width Non-Joiner and Zero Width Joiner, the sentence "The use of ZWJ/ZWNJ is not permitted in Internationalized Domain Names" is not accurate, at least not for IDNA 2008.

It could say "have been ruled out from the root zone by the [Procedure]".

# Comments not addressed from earlier versions

The next sections list comments from previous versions that the IP feels are still open (or partially unresolved). If the GP feels a particular comment was already resolved, the IP would like to ask for a short note identifying how this was addressed. Some issues may be unresolved in substance, even if there were some intervening changes to the text. Generally, the descriptions of the issues have not been updated to track the details of such intervening changes, so there may be some small discrepancies.

## General Comments

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| Overall | A large number of smaller editorial suggestions were supplied in the document …-IP which is a marked-up version of the LGR proposal of Bangla-LGR-20180626-IP.  These issues were too numerous to propagate forward to the latest review document, therefore the IP is sending ***both documents.*** For additional changes see Bangla-LGR-20180809-IP. | Please review this older document carefully and correct anything that has been flagged in there. (View with tracked changes). |
| 3.3.6 Nukta | Section 3.3.6 "Nukta" is very confusing as it repeats the description of Nukta from the Devanagari proposal verbatim with the Bengali-relevant issues casually inserted. This needs to be cleaned up. | Remove text not specific to Bengali. |
| Joiner | Zero-width joiner and Zero-width non-joiner are not part of the MSR, because the Procedure rules out any CONTEXTJ code points.  However, these code points are commonly used in Indic scripts to control the formation of conjuncts. That means that not incorporating them into an LGR prevents certain terms from being displayed and/or represented correctly.  We do have in all NeoB LGRs a part (§ 3) that purports to give the background information for the script; these chapters go into great details on things like the early history, but are curiously silent on the joiner characters.  We understand there is a brief mention in section 7.1.1, but it is not solely focused on joiner characters. | Please add. |
|  |  |  |
| References | 1. For websites, please indicate the name of the Website (e.g. Omniglot) as well as page title in addition to the URL. As appropriate, prove an “accessed on” date. If a document has an identified author, publishing date, or serial number, please list these, even if that information is somehow also encoded in the URL. 2. [108] this seems to be incomplete – same issue in XML | Please make edits as suggested |

## Ya-phalaa and Ra-phala and Repha sequences

These sequences are special in that they allow some code points to appear in contexts not otherwise allowed by the rules. The IP proposes to the GP to merge the description of these from section 7 with the description in section 5, making new sub-sections “3.3.8 Ya-phala Sequences” and “3.3.9 Ra-phala and Ref Sequences”. (See editorial suggestions on pp 46-47 of the marked up document)

This would be the best place to describe how these sequences work in the context of the script.

As we will see, the ya-phala or VHCM sequences should be listed in the repertoire. They are currently listed in a separate table; that table can be merged with the main table or kept separate, as long as it is clear that these sequences are members of the repertoire. Also: Add the sequences to the XML file.

If that is done some of the context rules that become unnecessary.

The following two rules can be deleted from the XML. V1 H C1 M1 will be allowed because all valid combinations will be listed as a sequence in the repertoire (and context rules are not evaluated inside a sequence).

|  |  |  |
| --- | --- | --- |
| precedes-V1-and-follows-C1-and-M1 | ([:[**V1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_V1):])←⚓→([:[**C1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_C1):][:[**M1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_M1):]) | Section 7, Special case of Halant from WLE 7: H must precede V1 and follow C1 and M1 |
| precedes-only-C-or-N-or-precedes-specific-V-and-follows-C1-and-M1 | (:[follows-only-C-or-N](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#rule_follows-only-C-or-N):)|(:[precedes-V1-and-follows-C1-and-M1](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#rule_precedes-V1-and-follows-C1-and-M1):) | Section 7, From WLE 2 and 7, H: must be preceded by C or N, or preceded by V1 and followed by C1 and M1 |

For the context required for the Khanda Ta, we note that “S” ends in M1 (which is a subset of “M”). Therefore, term (:S:) can be removed without any change (and rule “S” can be deleted entirely).

Rule “P” ensures that a H cannot precede a Z unless it is following a C2. This rule can be retained, and term (:P:) can likewise be retained.

|  |  |  |
| --- | --- | --- |
| S | [:[**V1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_V1):][:[**H**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_H):][:[**C1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_C1):][:[**M1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_M1):] | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P |
| P | [:[**C2**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_C2):][:[**H**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_H):] | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P |
| follows-only-V-C-N-M-D-B-X-S-or-P | ([:[**V**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_V):]|[:[**C**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_C):]|[:[**N**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_N):]|[:[**M**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_M):]|[:[**D**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_D):]|[:[**B**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_B):]|[:[**X**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_X):]|(:[S](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#rule_S):)|(:[P](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#rule_P):))←⚓ | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P |

The ya-phala sequences begin with a “V1”, therefore they need a context rule that is appropriate for their first code point, which is the same rule as for V (as “not-when” rule):

|  |  |  |
| --- | --- | --- |
| preceded-by-H | ([:[**H**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_H):])←⚓ | Section 7, WLE 8: V: cannot be preceded by H |

M normally does not have a right-hand-side context, so the ya-phala sequences, which end in “M1” also need no right-hand-side context.

The Ra-phala is a C H C2 sequence and thus does not require special treatment in the WLE rules. The Ref is a C2 H C sequence, which also needs no special treatment, other than allowing a “Z” (Khanda Ta) in place of the trailing “C”.

Rule P for the Ra-Halant can be kept in Section 7 and the XML, but all the description text should move to section 3 (as suggested above).

In rule “7” the term “S” can be deleted as can the <rule by-ref=”S” /> in the XML. Also, note to delete the “S” in the names of any rules in the XML. The context rule for 09CD needs to be changed to “follows-C-or-N” and these two rules “precedes-V1-and-follows-C1-and-M1” and “precedes-only-C-or-N-or-precedes-specific-V-and-follows-C1-and-M1” deleted.

Summary:

1. Create two subsections in section 3 and collect all description of the linguistic function of the sequences there.
2. Keep table of ya-phala sequences and mark them as members of the repertoire
3. Add them to the XML with context rule same as for V
4. Delete rule “S” from section 7 and XML; delete all references to “S” from rules and names.
5. Keep rule “P”.
6. Delete two rules in XML related to halant (now taken care of by sequence) and change context rule for 09CD to “follows-only-C-or-N”.

## Variants and WLE Rules

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| Missing Variant | **Visarga:** the Visarga would seem to make a candidate for variant between Bengali and Devanagari (it's not used in Gurmukhi). In both, it is relatively unconstrained and therefore could appear in whole-script variant labels made up from other variant code points. | Please investigate the Visarga |
| Section 7 | This section details 7 restrictions which apply to “Bengali“ (apparently, the script when used to write Bangla, but not necessarily for other languages). As such, they should not figure in the Root Zone, at least not without linguistic evidence of their widespread validity.  Item 5 appears to require use of Avagraha 09BD. But this is excluded from MSR. Presumably, therefore the “permissible combinations” are not permissible in the root zone. (It is fine to document such features of the script, as long as the relation to the Root Zone restriction and/or Root Zone LGR are made clear.)  The status of items 6 is unclear. The constraints mentioned do not appear in the LGR; are they simply for information or are they supposed to be added somehow to the proposal (.docx, and.xml)? | Please clarify status |

## XML file

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| **<description>** | 1. Because of some sensitivities about the naming of the script, an explanatory sentence is suggested mentioning Assamese in particular as one of the languages covered by this LGR. 2. The section on consonants can be streamlined by removing the details about phonetic classification. Such classification, while traditional, does not affect the way consonants are used in IDNs or in this LGR. | Please note suggested edits. |
| **<references>** | online references are bare URLs, need Website name, Page title, and, if identified authors/editors, plus “accessed on” date. Reference 108 appears incomplete | Please complete references |
| **Comments on <char>** | Generally, the IP like to see the language information in the comments. That information cannot be looked up elsewhere in creating the HTML files, unlike character names. The latter can be omitted.( If its desirable to have both, put character name first, then use “:” as separator) | Please provide the language information from the table in Section 5.1 in the comments for each code point. |
| **<rules>** | **WLE rules:** In the following set of rules defined to constrain the Khanda Ta (Z), the rule "S" is ineffective, because any [:M:] already matches the left-hand context for Z whether or not is also satisfies rule "S".   |  |  |  | | --- | --- | --- | | S | [:**V1**:][:**H**:][:**C1**:][:**M1**:] | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P | | P | [:**C2**:][:**H**:] | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P | | follows-only-V-C-N-M-D-B-X-S-or-P | ([:**V**:]|[:**C**:]|[:**N**:]|[:**M**:]| [:**D**:]|[:**B**:]|[:**X**:]|(:S:)|(:P:))←⚓ | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P |   These rules are affected by the recommendation to list ya-phala sequences explicitly as members of the repertoire. (See discussion above) |  |
|  |  |  |
| **Sequences** | The attached XML shows how sequences could be implemented and how the rules would have to be adjusted. | Please review. |
| **Detailed editing** | Copy of XML included with suggested changes. | Please compare to the version submitted for feedback and note suggested changes, review and use as basis for further edits. |
|  |  |  |

## Test Files

NOTE added Aug 14th 2018: Older issues related to the Test Files are most likely now moot, as the test files, unlike the other files, have been revised extensively.