Response of IP to Devanagari LGR Draft of 20th May 2018

DATE: 2018-05-29

# Overview

This document provides IP response to the Devanagari LGR proposal dated 2018-05-20, including XML and test file.

# General Comments

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| Current situation | The IP notes the work done by the GP in response to the previous feedback.  The IP has noted a possible serious technical issue with some of the proposed variant sequences. The IP is investigating that issue and not ready to make a definite recommendation at this time.  However, the remainder of the Devanagari LGR proposal is nearing finalization.  A few must-fix issues were found:   1. a small regression in WLE rules in the XML 2. a small regression in ref numbering in the XML 3. some discrepancies in the annotations (languages, references) between table 6 and the XML   These are detailed below.  In addition, there are a few minor editorial issues that could be taken care of. | Until the IP is able to give recommendation how to resolve the technical issue, the IP recommends, the proposal should not be put to public comment.  However, the comments and other feedback in this document are unrelated, therefore we encourage the GP to review them and make suggested fixes in the meantime.  Once the IP has a definite recommendation on the variant issue, we will communicate that separately. |
| Newar/ Newari | Both spellings occur in the document | Please review |
| Minor editorial | Some minor editorial issues are flagged in the change- tracked document. | Please review |
| § 5, Table of repetoire | Please review the example languages and reference against the data in the XML file (for convenience we are including an HMTL file showing how the data would be formatted in the published RZ-LGR).  In the IP’s view, It is acceptable for the XML to list all languages for which references have been given for a certain code point, while the document can continue to use the “Most languages given in section 3.2” wording).  GP should consider adding the references as needed, though. | Please review |
| § 5, Table of repetoire | In particular:   1. The references column has no entries for Nepali [113], Konkani [112], Bodo [110] or Maithili [111] even though these languages are mentioned. 2. The language column has no entry for Kashmiri, even though ref [108] is listed. 3. There are some differences between the references listed in this table and the references given in the XML file for the same entries. | Please edit |
| § 5.1 | Section 5.1 Devanagari section of MSR-3  ‘Not PVALID in IDNA2008, or are ineligible for the root zone (digits, hyphen) - White background’    This is not correct anymore, in MSR-3 white only defines not PVALID, the ‘ineligible for root zone are now pinkish’, not white. MSR-3 states:  “Code points … which are PVALID in IDNA2008 but excluded from the MSR for various reasons are shown with [a] pinkish  [background] …. Code points shown with a white background are not PVALID in IDNA2008.” | Please fix |
| § 5.4.4.3 | There is a slight discrepancy in the description around the maximum number of consecutive consonants:  Section 3.3.3 says that you can have in practice up to 5 consonants, but the limit is not enforced, so far so good.  Then section 5.5.4 3) refers to a sequence of consonants (up to 4).  Then as expected section 7 (WLE) does not enforce that number of consecutive consonants.    Perhaps section 5.5.4 3) could also say something about the limit not being implemented in the rules and address the apparent inconsistency between 4 and 5.  Note that the XML description, while cursory, does not have any issues: “The Halant thus joins two consonants and creates conjuncts, which can be generally from 2 to 4 consonant combinations. In rare cases**[,]** it can join up to 5 consonants. However, this LGR will not enforce any length limit.” (there is a comma, missing, though) | Please fix |

# Whole Label Evaluation Rules

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
|  | At this point, the IP has no issues with the content or formulation of the WLE and context rules. (See report of test results below). |  |

# XML file specific

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| XML: <rules> | Two class definitions had a typo (found by checking the test file):  <class name="V1" from-tag="Matra" …  <class name="M1" from-tag="Matra" …  should have been  <class name="V1" from-tag="V1" …  <class name="M1" from-tag="M1" …  Note, the Tag values V1 and M1 were assigned to the correct code points, but the classes were referencing the wrong tag value. | Please make sure to use the fix from accompanying XML file. |
| XML: <description> | No issues at this time |  |
| XML: <data> | 1. The new reference for Nepali changed the numbering for the reference for Konkani, invalidating all entries “112” (in the accompanying XML these have all been changed to “113”). 2. None of the Nepali code points had been annotated with the new reference (see XML). 3. in-script sequences were inconsistently given or not given ref=”0” 4. The references listed do not agree with those shown in table 6 in the main document in some cases 5. The language names may not agree in all cases with table 5, however, instead of using “most languages in section 3.2” the XML just lists all languages for which a reference is given (that kind of difference is OK) 6. Kashmiri is usually documented with references 105 and 108 (not an issue) but sometimes only one of them – is there a reason? | Please review that the references and language names in the comments now match what is expected. |
| HTML version | The IP is attaching an HTML file formatted in a style close to what will be used for “Element” LGRs in the published Root Zone LGR. The purpose of sharing this is to all the GP to more easily compare the contents of table 5 with the data in the XML. | Please use this file in verifying the ref and comment attributes against table 6. |
| **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

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| Test file | * 1. initial testing found one small bug in the XML (regression from recent edits)   2. all valid/invalid dispositions agree   3. there are a large number of duplicate labels, these do not contribute to the test results   4. in some cases the duplications arise since test labels are divided by language, in those cases the duplications are unavoidable and document how languages overlap;   5. however, in sections of the test file not divided by language, duplicated labels just make it more difficult to quickly verify the test results. | Generally the test file was well organized and to the point; the results agreed (as far as the tools allowed direct comparison) |

## Test file analysis results

The IP evaluated both the test file delivered, “Devanagari-test-labels-20180520.txt” as well as a separate corpus.

Labels where the disposition in the test file disagreed with the disposition determined by the IP’s LGR toolset.

Initially we picked up some cases like this one:

**#Invalid - Case of "N: must be preceded only by either of specific set of Cs, Vs and Ms"**  
Validation: ‎कौ़ल‎ (0915 094C 093C 0932) : valid

which we realized was due to a small bug in the WLEs (see section 4 above). After fixing this in a revised XML, the test file and the tool now agree on valid/invalid status (except that our tool reports duplicated or repeated labels as such and not with a disposition).

**#Following labels produce cross-script variants with Bengali  
#However, as these would be invalid (partly Devanagari and Bengali) variants as the entire strings are** **not composed of variant characters/combinations**Validation: ‎महाराष्ट्र‎ (092E 0939 093E 0930 093E 0937 094D 091F 094D 0930) : valid  
Validation: ‎किताब‎ (0915 093F 0924 093E 092C) : valid  
Validation: ‎मुखपृष्ठ‎ (092E 0941 0916 092A 0943 0937 094D 0920) : valid  
Validation: ‎आदि‎ (0906 0926 093F) : valid  
Validation: ‎अंतिम‎ (0905 0902 0924 093F 092E) : valid  
Validation: ‎जिमखाना‎ (091C 093F 092E 0916 093E 0928 093E) : valid  
**#Following labels produce valid cross-script variants with Gurmukhi**Validation: ‎घंटी‎ (0918 0902 091F 0940) : valid  
Validation: ‎ठेट‎ (0920 0947 091F) : valid  
Validation: ‎भिम‎ (092D 093F 092E) : valid  
**#Following labels produce valid cross-script variants with Bengali**Validation: ‎ममममम‎ (092E 092E 092E 092E 092E) : valid  
Validation: ‎मिम‎ (092E 093F 092E) : valid

Looking at the code point sequence for all of these it looks like they are all Devanagari-only labels, so the original labels should be valid (even if some variant labels are not valid labels in either the Deva or Guru LGR).

Evaluating the variant labels required splitting the file, because the tests performed are different. Running the 1262 labels from the sections that identify labels with variants we found the following:

Total Labels processed: 1262 of which

valid labels: 755

has variant: 755

invalid labels: 0

skipped labels: 507 of which

duplicate labels: 507

Even after separating the valid/invalid part of the file there are still 507 labels that are simply duplicates of another label in the same file. Those could be removed; they do not add any value in testing.

Of the remaining 755 labels, all 755 were reported as having some variant. For “blocked” variants, we do not generate the set of possible variants with this tool, so we cannot evaluate which labels can form valid variants in another script.

The following 13 labels were found to have matching variant labels inside the original file. They are shown here with the “Index variant” (common variant label with lowest code point values in all positions).

‎आले‎ (0906 0932 0947) <==> ‎आलॆ‎ (0906 0932 0946): Index Variant = ‎आलॆ‎ (0906 0932 0946).

‎आले‎ (0906 0932 0947) <==> ‎आलॆ‎ (0906 0932 0946): Index Variant = ‎आलॆ‎ (0906 0932 0946).

‎आले‎ (0906 0932 0947) <==> ‎आलॆ‎ (0906 0932 0946): Index Variant = ‎आलॆ‎ (0906 0932 0946).

‎ज़यादॆ‎ (091C 093C 092F 093E 0926 0946) <==> ‎ज़यादे‎ (091C 093C 092F 093E 0926 0947): Index Variant = ‎ज़यादॆ‎ (091C 093C 092F 093E 0926 0946).

‎ज़यादॆ‎ (091C 093C 092F 093E 0926 0946) <==> ‎ज़यादे‎ (091C 093C 092F 093E 0926 0947): Index Variant = ‎ज़यादॆ‎ (091C 093C 092F 093E 0926 0946).

‎ज़यादॆ‎ (091C 093C 092F 093E 0926 0946) <==> ‎ज़यादे‎ (091C 093C 092F 093E 0926 0947): Index Variant = ‎ज़यादॆ‎ (091C 093C 092F 093E 0926 0946).

‎रोहो़ड़‎ (0930 094B 0939 094B 093C 0921 093C) <==> ‎रोहोड़‎ (0930 094B 0939 094B 0921 093C): Index Variant = ‎रोहोड़‎ (0930 094B 0939 094B 0921 093C).

‎रोहो़ड़‎ (0930 094B 0939 094B 093C 0921 093C) <==> ‎रोहोड़‎ (0930 094B 0939 094B 0921 093C): Index Variant = ‎रोहोड़‎ (0930 094B 0939 094B 0921 093C).

‎रोहो़ड़‎ (0930 094B 0939 094B 093C 0921 093C) <==> ‎रोहोड़‎ (0930 094B 0939 094B 0921 093C): Index Variant = ‎रोहोड़‎ (0930 094B 0939 094B 0921 093C).

‎हा़जिर‎ (0939 093E 093C 091C 093F 0930) <==> ‎हाजिर‎ (0939 093E 091C 093F 0930): Index Variant = ‎हाजिर‎ (0939 093E 091C 093F 0930).

‎हा़जिर‎ (0939 093E 093C 091C 093F 0930) <==> ‎हाजिर‎ (0939 093E 091C 093F 0930): Index Variant = ‎हाजिर‎ (0939 093E 091C 093F 0930).

‎हा़जिर‎ (0939 093E 093C 091C 093F 0930) <==> ‎हाजिर‎ (0939 093E 091C 093F 0930): Index Variant = ‎हाजिर‎ (0939 093E 091C 093F 0930).

‎हा़जिर‎ (0939 093E 093C 091C 093F 0930) <==> ‎हाजिर‎ (0939 093E 091C 093F 0930): Index Variant = ‎हाजिर‎ (0939 093E 091C 093F 0930).

For the entire file before separating out the variant related part

Total Labels processed: 3173 of which

valid labels: 1163

has variant: 1105

invalid labels: 48

skipped labels: 1962 of which

duplicate labels: 1950

broken labels: 9

contain join controls: 3

start w/ wrong script: 0

2/3 of the labels are duplicated ad therefore ignored by our tool, however about 350 additional labels from the valid/invalid part of the file were also found to be having variants.