Suggestions for LGR

I have gone through the supporting documents for LGR Proposal. I have not just learnt about Indic computing from the basics of the principles followed in the start, but also seen through the complete evolution of standards to this time. I must congratulate this “particular” effort whole heartedly. The document is not just very well done, but also covers the most useful aspects. Though it is guided by the need for LGR, this seems to be the first such work “after” the initial standardization work for Indian languages done nearly four decades ago. The authors and contributors deserve due honor.

Given the above, I will urge the effort be extended not just to LGR for Indic computing in general. Since the panel exists as such a useful and active group right now, extending the work seriously across Indic computing will make it a lot more useful and serve the purpose for which the effort on IDN in Indian languages has been started.

Below are some of the inputs I observe. I am commenting only on section 3 which is the actual recommending section and not getting into the rest.

1. In 3.3.2, the additional mention “after” the first three sentences, are not important. It may add needless confusion.
2. 3.3.3 mentions mentions two vowels U+090D/ U+0972 as same with a note as Marathi uses ॲ (U+0972) instead of ऍ (U+090D). This violates the unambiguity guideline for LGR. Also, this is a significant deviation for the nature of the script definition and is also a reason that generates strong debates why Assamese and Bangla are to be considered different scripts. The fundamental reason is that Indic scripts being phonetic, there are no two characters in a script that will be identified as representing the same identity in any language using the same script. In this particular case, the above confusion carries no basis. The vowel in context here has been created for phonetic representation of pronunciations for English words. These vowels are not taught in the native alphabet in schools. Since the character definition has never officially evolved (except perhaps in encodings), the confusion mentioned here is only a debate in pockets. Students of these languages have not natively been taught and a standardisation can actually remove this confusion and have students learn only one form and the usage will also have a uniform practice. This will also remove ambiguity within the script. In fact, seeing through a lot of Hindi corpus (the language perhaps most used with Devanagari script) does not show the use of ॲ (U+0972) or ऍ (U+090D). The matra form is seen for some English words though. Perhaps the character finds two encodings within Unicode due to lack of clarity and that standardising the visual form for only one encoding will be the answer. LGR may use only ऍ (U+090D) and state that the form used could be ॲ since the use of ऍ is not seen. In fact. The character ऍ was encoded in ISCII (The first encoding standard for Indic scripts) and the experts who developed the standard, had Marathi experts who actually introduced the vowels ऍ and ऑ and hadn’t suggested a ॲ. More than 3 decades ago, ऍ was not used in any mainstream Hindi (either officially or unofficially) and English words needing the pronunciation for words like “bat” and “ball”, wrote (and still do in most places) as बैट and बाल.
3. 3.3.4 may need more definition at least for normalisation. The variant guideline clearly demands that if two valid forms may exist, they will have to be normalised. While the definition in this section states the use of Anusvara with respect to the varga consonants, it doesn’t define the same for the rest. Will कंस, कन्स and कम्स be variants of the same or different?
4. 3.3.5 mentions that present day Hindi users tend to replace this with anusvara. I am not sure if this is a factor. The different between हँस (laugh) and हंस(swan) is not disambiguated by such a rule. I suggest the second sentence be removed. People may make common errors in spellings and that happens in every language but that must not become a norm.
5. My personal suggestion is to disallow Nukta. Nukta is “not” a character nor does it create or identify a new character. Nukta is an accent marker. There are no two words that differ only by the nukta and have the same meaning. The word nukta itself was borrowed from Urdu where it is used as an accent marker. In Indian languages, wherever it is used, it only appeared under ड and ढ wherever the pronunciation for these consonants were flowing than stern. But, the nukta never participates in any conjunct. Use of this in regular text is for a visual representation for accent marking but is ignored in computing and words written with or without nukta carry the same status. Considering the unambiguity need in IDNs, the use of Nukta may not be permitted. But, if the committee experts decide to permit use of Nukta, then the indiscriminate joining of Nukta with any consonant or vowel may not be allowed. Only the valid lists may be outlined.
6. I am very pleased that NBGP experts have discouraged the use of ZWJ.

It does appear that the use of ZWNJ cannot be avoided. MSR doesn’t permit the use of ZWNJ. But, this is due to a fundamental issue with encoding. As defined in 3.3.2, Halant is the implicit vowel (schwa) remover. But, Halant is “not” a joiner. Hence, most of the Indian languages have many words that are written with halant but do not end up forming conjuncts. To form conjuncts, students of these languages learn about joining (yukta). Ideally in encoding, these two characters should have had such nomenclature so that Indian users can associate with the way they have learnt their languages. I am not sure if NBGP has the scope to influence this, but looking at the work, my expectation is certainly towards the same. Halant must not behave as a joiner and must remain visually explicity. For joining, the joiner may be used and the character be called as Yukta. This suggestion should ideally go to the Unicode consortium, but since this also relates to the behavior of halanta, I am not in a position to make it. ZWJ and ZWNJ are not characters nor are character operators. Hence, use or non-use of these are unclear and debated.

1. In section 5.2, what is the difference between U0912 and U094A? U094A is mis-represented. Should be the matra form.
2. Section 5.2, the note for table 7 refers to footnote 13. The footnote explains that the use of U0931 is only for a display variant. This is in line with the actual language use. The languages Marathi, Konkani and Nepali do not have a ra-dot (RRA) as in U0931 and hence, is not taught as a part of the language. However, the conjunct with some consonants do take the display form of eyelash reph. It is a legacy from ISCII where technology was limiting the computing systems of those times to have different display forms in different languages for the same conjuncts. This is not different from the alternate conjunct forms and lead to ambiguity in use. To remove ambiguity, the use of RRA may not be permitted. In case the panel experts decide otherwise, then it must be stated clearly that use of RRA is only for a display variant so that IDN normalizing for search etc. can follow the same guiudeline.
3. Section 5. I suggested that use of nukta be avoided. However, if that be permitted, the list of Nukta consonants must be mentioned in this section (5.5.4) to make it explicit. Also, I presume this section only attempts to mention the validity of character sequence. Though such validity “is” dependent on the akshar definition, but emphatic mention about akshar, number of consonants that can join etc. are perhaps extra information and will needlessly raise confusion and questions. These definitely have a very important need if the document will also be referred for display implementation (fonts and fonts rendering). Unless such is an intention, the additional information may be removed.
4. The sections after that reiterate or elaborate about halant and nukta which will probably get revised only based on adoption of the above suggestions.

I will re-emphasise my appreciation to this work and suggest that this be extended to Indic computing in general. This guidelines for LGR are not necessarily minimalist but are very practical and if extended to Indic computing in general, will make the growth of Indian languages use, easier and faster. If the NBGP panel of experts may agree to shoulder such a responsibility to take this amazing effort beyond LGR, then this may also include font definition formats. The most prevalent font format today is OpenType which has very complex tables and font designers find it very hard to understand and create new fonts. This is supported by the fact that the number of fonts and designs that grew before OpenType became the only supported format on Windows and later other OS, dropped drastically to very few options. Language use and publishing will not grow without good fonts and development of fonts is restricted by the complicated font format and lack of definitions and standard. Also, the OpenType definition works over glyphs only. Hence, the character classification as outlined in this document in section 3 and script grammar as outlined in section 5.5 are critical parts to the display behavior for Indic scripts and hence, OpenType rendering engines pose limitations.