EN

ABU DHABI – IDN Neo-Brāhmī Generation Panel Meeting
Tuesday, October 31, 2017 –13:45 to 15:00 GST
ICANN60 | Abu Dhabi, United Arab Emirates

UNIDENTIFIED MALE: It is October 31st 2017, IDN Neo-Brāhmī Generation Panel Meeting, Capital Suite 6, 1:45 in the afternoon.

SARMAD HUSSEIN: Welcome, everybody, to the Neo-Brāhmī Generation Panel Update to the Community at ICANN60. Let me start by handing over the session to the Neo-Brāhmī Generation Panel co-Chair, Dr. Ajay Data to take the lead of the panel.

AJAY DATA: Thank you, Sarmad, and welcome, everyone, for the Neo-Brāhmī Generation Panel Update. This is basically a panel based out of India right now but covering some countries beyond India which are Nepal, Sri Lanka, and Bangladesh and [inaudible] because of the common scripts use.

We have been active for the last few years and the way we have been functioning is to speed up things as we are having fortnightly calls and review of the documents through NBGP mailing list.

Note: The following is the output resulting from transcribing an audio file into a word/text document. Although the transcription is largely accurate, in some cases may be incomplete or inaccurate due to inaudible passages and grammatical corrections. It is posted as an aid to the original audio file, but should not be treated as an authoritative record.

We have around 40 committee members actively participating to that and we are able to submit NBGP Devanagari proposal, which is one of the scripts out of nine total scripts under [inaudible] for IP. And a few days before we got our first review by IP and we have got some suggestions to be incorporated and we could not look at that in detail because after that ICANN started and we all are here so as soon as we are back, we will look into this IP suggestions and incorporate those changes or take up the decision on them.

We have also gone to Nepal to include the community from there who are Nepali and many other languages which are common for Devanagari. We are also going to Sri Lanka in December for Tamil and also cover Sinhala which is the language in Sri Lanka, and we're already planning for Bangladesh in March to cover Bangla and include more better participation.

This is the our first face-to-face meeting though these members are already in the mailing list and active and whatever suggestions they are doing right now are being considered participation. They're also participating on online calls through remote participation and I think we have done reasonable progress on Devanagari and we have taken some very strong commitment to the morning call of the Community of NBGP Meeting we had today. And we are going to finish another eight

very soon though we are working with [inaudible] teams, now with eight since we have a template ready, we have the structure ready, we know what it takes to do that and we are going to now. We're panel eight LGRs together and have dedicated teams and try to work with all of them together.
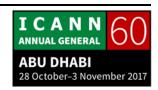
So that's, I think, a small brief from NBGP. Anything I missed, Sarmad? Now I will let Akshat present the document and share with all of you what we have gone through on Devanagari. Thank you very much.

AKSHAT JOSHI: Thank you Dr. Data. I'll be presenting the Devanagari NGR at a later part of this presentation. Before that, I would just like to take everyone through some of the nitty-gritties or all the scripts that come under [and bit] of Neo-Brāhmī and then we'll specifically be diving into the Devanagari part of it.

We are Neo-Brāhmī Generation Panel co-chaired by Dr. Ajay Data, Mr. Mahesh Kulkarni and Dr. Udaya Narayana Singh. Why the name Neo-Brāhmī is one of the questions that can be there in the minds of community all the time.

When this project IDN Variant TLDs project started, ICANN had already identified six script families which have their unique set of way of representations. Under that Devanagari was

considered as a representative case for this kind of language families and we participated into the process and there are many inputs which went into the formulation of the process in a particular way that could cater to the requirements of all the languages which fall under the Neo-Brāhmī family, specifically in terms of whole [label] evaluation rules.

However, having come out of the process, we knew that it's not only about Devanagari, it's about the whole of our language scripts that have been derived from the Brāhmī family that need to be taken care of in this particular way in which we had envisioned during the process formulation.

So the decision was taken that let us make an umbrella panel which will take under its cover majorly nine scripts, as Dr. Data just spoke about now. And the effort of the panel is always that the nature of decisions that we take across the board should be more or less similar for each of the scripts.

So when we say Brāhmī family scripts, what are the basic features of these scripts? They're written from left to right and there are certain categories of characters. Those are used in those scripts. One is a consonant, the other one is a vowel. A consonant and vowel is something which is easily known by everyone.

Apart from that, we have certain vowel modifiers which sit on top of consonants and modify the sound of it. Then we have certain characters which modify both consonants as well as vowels. They are called [inaudible] chandrabindu [inaudible]. Then we have a peculiar character known as halanta, which comes in between two consonants and creates a [inaudible] out of that so that is also a particular requirement.

There is a specific format to which characters can come after one another and this was given by one of the Indian Standards ASCII 91, the way it has been already called, and this order we want to preserve even in the Label Generation Rule Sets.

Why I am saying this is for example there is a case where a consonant can take a vowel modifier as well as a chandrabindu or [visarga]. Now confusion can always arise whether the chandrabindu comes first or a vowel modifier Matra comes first, and that has already been resolved by a very structured way and that is already there in the formalism.

So in the next part of this presentation, I'll take you briefly through those constructs. I don't want to be very much technical about it but I just want everyone to know the kind of nitty-gritties that are involved and then we will go to the way we have framed the Whole Label Evaluation rules, which are quite

ICANN 60
ANNUAL GENERAL
ABU DHABI
28 October–3 November 2017

EN

different from what I'll be showing in the middle of this presentation as a requirement for Brāhmī scripts.

This comes out of the fact that the rules that need to be designed for Label Generation Rule Sets. One of the key principles is the simplicity principle which demands that all the rules that we frame should be quite simple to understand. So we'll go in details in this way.

So what are the scripts that come under the Neo-Brāhmī umbrella? There is a lot of text available. I'm sure you can refer to the slides afterwards and go through that.

I'll just go through the names of the script. One is Devanagari, the other one is Gujarati, Gurumukhi, Bengali, Odia, Tamil, Kannada and Telugu, and Malayalam. Apart from that, Sinhala script also kind of comes in this family.

However, at the time of formation of the panel, we were not sure whether we will get the right expertise for taking this work further so we did not initially include Sinhala into it but with the recent changes where we are now going to Sri Lanka where there is predominantly Tamil as well as Sinhalese community residing, we have two motivations of going there. One is getting on board the Tamil community in Sri Lanka as well as establish coordination with the Sinhala Generation Panel and ensure that

we are in sync with each other going by the larger umbrella principal of the Neo-Brāhmī.

So, current status of the work is that Devanagari LGR is almost ready. It has gone through two interim reviews with the integration panel and the next submission that we will be doing soon. It will be a finance submission as of the Devanagari LGR from the Neo-Brāhmī Generation Panel.

The idea or method of working for this was that once we have the Devanagari LGR in place, we have a template ready for all the other scripts to follow and that was envisioned to speed up the work for every other script.

We have not per se started on the document. However, the basic components of other scripts are already being developed by the community members. So we are in good shape to… once we are done with Devanagari we will immediately come up with next set of LGRs.

Before going to the Devanagari LGR, I would like to take everyone briefly through the actual formalism that binds the Brāhmī-based script.

As I discussed earlier, these are the basic components and here we can see consonants and vowels then there are Matra signs which are called vowel signs and you can see there is a dotted

circle on that. That is an indicator that this character stand alone cannot exist unless it sits on a particular consonant character.

Then there is a set of vowel modifier characters which can sit both on consonants as well as vowels. They are also shown by the dotted circle for the same reason. Then there is a Halanta character I discussed earlier which forms a conjunct between the two consonants, and then there is a peculiar character called a Nukta.

Our initial assumption was that it only sits on a consonant to modify its sound. However, during the language community outreach, we came across requirements of Santhali language which needs the Nukta character to sit on the vowel as well as vowel Matra so we had to considerably change our rule-set for accommodating that request. We'll see that at a later time.

So what is this formalism? I'll be brief about it. It is a bit much more technical and I don't want to explain each and every part of it because it might take a lot of time but I just want everyone to know the kind of background work that has gone into it.

These are the set of variables… these are much more than what I earlier showed as a basic set of characters. Why this is the case is because when we think about all the kinds of scripts that come under Neo-Brāhmī Generation Panel, each script sometimes uses certain characters which do not have any

parallels in the other script by way of their behavior. So when I'm saying this I mean the Gurumukhi has a particular character known as Addak which doesn't have any parallel in any other script so it forms its own separate category.

Then Malayalam has a Chillu Aksara which have their own formations. Bengali has a [jafala] construct which is again peculiar to that. So all those peculiar things have been considered under this formalism and a separate space was given to them in this formalism.

However, when we will be creating the LGRs, since those will be separate LGRs for each of the scripts, only that LGR which will be having these particular characters will be talking about this character. So this being a general formalism for all the Indians Brāhmī-derived script, this has found in this particular slide at a common place.

So these are just basic sets of operators which are used for value denoting the formalism. The entire formalism at a glance looks like this. There are two main components. One is consonant syllable, the other one is vowel syllable, and you can see there is quite a lot of things going on over here.

I would like to take your attention to the consonant syllable first. The first line after *k(C[N]H), there is a C. That is the base character and then everything else is optional so if you want to

look at it, these are the kinds of characters which can sit on a consonant. Then there is a previous construct, that k(C[N]H). That is something which can come before the consonant to form a conjunct.

Especially this construct was designed to cater to the need of different conjunct depths as employed by different languages using various scripts. For example in Hindi the conjunct depth generally doesn't go beyond three consonants whereas if we take a case of Sanskrit, it goes much more beyond. It's like four or five conjunct depth.

This construct was designed to take that requirement into consideration. However while formulating the rules for LGR, we are thinking that this is a bit too restrictive for LGR and we have kept it kind of relaxed owing to the simplicity requirement of the LGR. The first line is majorly applicable for all below three lines in the consonant syllable. First one is about, if I'm not wrong, Bengali where Z means Khanda Ta character. The second one is about L which is about a Chillu character in Malayalam. And the third one which starts with AC is about Addak character as required by Gurumukhi. So this is the consonant syllable.

Vowel syllable is quite straightforward and is across the board almost the same. I'll not go into the details of that. And then

these vowel syllables and consonant syllables together constitute a syllable which can in turn create an IDN label.

Going by details of consonant syllables, there are examples of these. For example, this slide shows – at the top of the slide there is a part marked in blue and what all characters can come, the manifestation of that has been given on this slide. Again I'll request everyone if you are more interested in the details of it you can refer to this slide afterwards.

And now this is the slide where the previous construct comes into picture where k(C[N] followed by H can be preceded to the consonant to form a conjunct and there are various depths of conjuncts you can see on this slide which are shown.

The next one happens to be Khanda Ta character, which is required by Bengali and it has a peculiar requirement that it can stand on itself. I'm not a Bengali expert but I know that it can stand on itself. It is pronounced as Tha whereas it can be optionally preceded by a consonant and a halanta for forming a [reph] kind of construct which is like a ra halanta Khanda Ta which has a particular sound into Bengali. This is a peculiar requirement hence this character has been separated out.

Going to the next one, this is the Chillu character as required by Malayalam community. Chillu character has a peculiar requirement that it can stand by itself or it can optionally be

followed by a halanta and a consonant, which again optionally can be followed by a DNH. And then there is a Matra. If it can come it can again be optionally be followed by a D. so there are these cases listed here.

Then coming to the last one, the Addak character, the Addak character cannot stand by itself. It has to be followed by a consonant. AC seems to be the construct over there and then there are some optional things which can follow an Addak character.

Vowel syllable is quite straightforward. A vowel can stand by itself or be optionally followed by various symbols that are being shown. I see some disagreement from Anivar. You're talking about the Chillu character. I guess we will talk about it in the discussion section.

So this forms a syllable and ultimately we get an IDN Label derived out of it. Now this is the basic complexity behind it. However, being part of the process we knew that this kind of complex construct we cannot specify. There is a typical need that we should define it in clear terms which is in consonance with the XML rules so we had opted a different way of specifying rules and we will see that afterwards.

Coming to the Devanagari LGR, there are three major components of the Devanagari LGR. One is the code point

repertoire, which takes into consideration almost 10 to 12 languages. Apart from that, we had expanded our language outreach to EGIDS scale 1-4, which necessitated that we take on board many more languages for which the expertise was not readily available.

In the recent period we have extended outreach to those languages also and we are happy that not many changes were required to the basic Devanagari LGR and those languages are also being considered. Then one of the very important factors of the Devanagari LGR is Whole Label Evaluation rules which have been specified.

The example being shown here is the word Kitab. Kitab means book in Hindi, so the first one is correct. The second one seems to be correct but is not because there is an additional Matra followed by a Matra. The third one seems to be outright wrong and is flagged as wrong because the rendering engine as is available on most of the operating systems recognizes this as a fault and then it shows it that way. However, in the case of the middle one, most of the rendering engines fail to identify this as an anomaly and don't show this and we want to prevent these kinds of cases getting into the domain names or top-level domain names.

Then there are variant rules. In case of Devanagari or Neo-Brāhmī scripts, owing to their design, there are not many variants which look exactly alike to each other. There are certain cases which get weeded out because of the IDNA disallowing and everything so there are hardly any exact looking variants for the Devanagari or Neo-Brāhmī but there are quite a few cases of similar-looking which are confusing to the user. While incorporating additional languages we ended up by forgetting our definition of confusingly similar variants for the Devanagari and in the next slides I'll cover that.

So these are the principles that were specified in the Root LGR procedure and I'll take you through our take on each of the principles and how we have tried to adhere to those principles while drafting the Devanagari LGR.

So the Longevity Principal says that panels are supposed to begin using the latest version of Unicode, but also to take into consideration the stability of the Unicode character properties. So the basic set of characters, majorly almost all the characters that the Devanagari LGR proposes to be included exist in the Unicode right from Unicode Version 1 so they have quite stabilized properties.

The latest characters that Devanagari LGR includes were last encoded into the Unicode 6.0 which were specifically for

Kashmiri and which is again way back in October 2010 so we expect that the Longevity Principle we are adhering to.

The Least Astonishment Principle says that it ensures that allocated code points included in the zone repertoire are useful as elements in unique identifiers to the extent that code point is confusing to the user population or can be used in surprising ways – whether to members of the original linguistic target community or, in the case of the root, to the members of other linguistic communities – use of the code point fails to adhere to the Least Astonishment Principle in that context.

So in case of Neo-Brāhmī or Devanagari, to be specific, there are certain characters which we have identified as a part of variant set which might cause confusion to the larger portion of the Devanagari user base because those are the kinds of construct which are peculiar to certain languages and those are the kinds of construct which if not included would mean that those languages are as good as ignored. So there was a fine balance that we needed to strike and we have tried to do that by introducing them as blockable variants which enable us to incorporate them as well as take out the possible confusion which might arise.

Inclusion Principle says that by default each and every character is excluded and we have to selectively take each an individual

character and we have adopted that procedure while finalizing the character set.

Simplicity Principle says that part of the point having the integration panel is that it performs the check of Simplicity Principle. The integration panel cannot possibly include experts in every language and script but the members must have general knowledge of Unicode, IDNA, DNS, or all of the above. If any member of the integration panel cannot understand the rationale for inclusion of some rule, then that member will not support the rule and it will not proceed. This is the purpose of the unanimity requirement for the integration panel.

NBGP has tried to frame all the rules it frames in a very much simplified way. I just wanted to give you a glimpse of in what forms they earlier existed to give a sense of how much we have tried to simplify the overall rules. And under the NBGP we do have community members who have general knowledge of Unicode IDNA, DNS and registry registrar operations. So we have done the rule drafting in that particular way.

Predictability Principle says that it is as good as a Simplicity Principle that if the integration panel does not immediately agree with the recommendations of the generation panel or if the members of the integration panel disagree with each other, that is a good reason to suppose that rule in question is not

really predictable. So we can just hope that this case does not arise when the evaluation happens of the Devanagari LGR.

Stability Principle is as good as Longevity Principle so I'll skip this one.

Conservatism Principle is that the proposal is consistent with the Conservatism Principle in two ways. First and most important, because the integration panel is supposed to reject anything that it does not positively think is safe, the Conservatism Principle is built into the integration panel's criteria.

Second, in the event of disagreement between the generation and integration panels, the proposed rule that is the subject of disagreement is automatically excluded from the Root Zone LGR. NBGP has tried to see that this case would not occur.

This is the chart of the Devanagari MSR. MSR is Maximum Starting Repertoire and here the characters marked in yellow are the ones which were available for the generation panel to be included into the LGR. The ones marked in pink are already debarred by the MSR. The ones marked in white are barred by the IDNA protocol.

So out of the available character sets, Devanagari LGR excludes this set of characters and in the recent feedback from the

integration panel we came to know a mistake in the code point repertoire of code 0944 character so that can possibly be a part of the set as well.

And then apart from the character sets that are already proposed, there is one character that is 0931. It has not been included as a character but it has been included as a sequence of characters because it comes with a contextual element associated and we're only permitted those particular contexts and not the standalone 0931 character.

So this slide talks about all the Whole Label Evaluation Rules recommendation from the Devanagari LGR. As you can see, there are a certain set of basic categories and the first rule states about the Nukta limitations. The Nukta character can only be coming after a certain set of characters. The first set of characters are consonants, the second ones are vowels, and the third ones are vowel modifiers also called Matras.

The vowels and vowel modifiers requirement is specific to Santhali and the same has been included in the variant set for avoiding the confusion for non-Santhali Devanagari users, and then there is a set of general restrictions which are to be put on to the Devanagari label.

So the case of variants with the Devanagari LGR is ideally the LGR procedure is designed in such a way that the homographic

variants which are also called confusingly similar looking variants should not be part of the LGRs. There is a separate process which can take care of this. Initially when we were working on this, even we kind of were in agreement with this.

However, with the recent changes which necessitated us to incorporate certain characters which majority portion of the Devanagari community may not identify with, yet they're required for a certain language community. We had to introduce those set of characters and thereby variants.

So the two categories that we are defining of variants under Devanagari are pure homographs which are purely confusable because of the confusability of the shape. The second one R perceptive homographs where majority of the people do not expect certain characters to come in those positions and hence they might get confused. And as we see it's a serious kind of confusion because they might think that that is the same thing, it's just a minor font variation, and we wanted to avoid that from happening.

Dr. Samuel, we're on time, right? Because I have still ten more slides to go. Okay, this slide shows the first category of variants which is purely confusingly similar-looking. You can see merely by looking at the characters there seems to be some resemblance but one can immediately make out the difference.

However, when we are talking about domain names or top-level domains, if we see the use case, they're mostly shown into the browser URL bar and at that point if we see some of those characters and if we analyze them vis-à-vis the various default forms that various operating systems and browser combinations employ on them, these can come under confusingly similar case. So these have been depicted as such but not proposed as a mandatory set of variants under Devanagari.

So these set of variants is actually a Nukta set of variants which are due to the Santhali requirement of having Nukta after a vowel, any vowel signs. What is Nukta? Nukta is just a small dot which sits on a vowel or a vowel sign and again looking at the browser point size and everything, for a user that doesn't know that this Nukta –because generally Devanagari is very familiar with a Nukta kind of character coming below a consonant. We expect that this could happen and hence we make the distinction, whereas non-Santhali users will never expect this Nukta to come below a vowel or a vowel sign and it is so subtle that they might just miss it. So we want to ensure that this is permitted and the other set of labels which can get generated are generating blockable variants.

This slide shows some of the characters – I'm sorry about the rendering but some of the characters are not being rendered properly. But these are the Kashmiri characters which we have

included and Kashmiri language absolutely needs these characters for Kashmiri representation in Devanagari. However, their nature is as such that they look like certain other sets of combinations and we are proposing them as variants as well.

Then there is an interesting case of cross-script variants. When we are talking about these, these cross-script variants are not within the label but there is always a possibility that an entire Devanagari label can probably look like an entire Gujarati label or Gurumukhi label. I'm not saying that they do but there is always this possibility and at a larger level which goes beyond the scope of a particular LGR. These are also potential cases which need to be considered so we're proposing that every LGR will have an analysis of possible cross-script variants and here we have identified some of those cases. But this work is ongoing and we continue to enhance this [chart] as we go ahead.

So coming to the methodology, we recently had a face-to-face meeting in May in Kathmandu, Nepal, and we got a very good active participation from a community from Nepal as well as for the Newar and Tamang languages. We got very good feedback on that. And we have the third face-to-face meeting scheduled to be held Sri Lanka which will enable us for better understanding of Tamil as well as coordination with the Sinhalese Generation Panel.

Then in terms of language coverage, we had considered that we will take into consideration the languages falling within EGIDS scale 1-4. In the very initial stages of the Devanagari LGR, we had taken into consideration Bodo, Dogri, Hindi, Kashmiri, Konkani, Marathi, Maithili, Nepali, Sanskrit, Santhali, and Sindhi languages. However, recently with very active outreach effort we could cover Avadhi, Bhatri, Bhojpuri, Chhattisghari, Halbi, Kinnauri, Kukna, Limbu, Magahi, Newar, Panchpargania, Sadri, Tamang Eastern, and Wagdi. These are the languages which are actually being spoken by a language community which is sometimes very remotely located and it was quite difficult to get them to understand the concept of what we are doing and getting their feedback but we could get that done. Jay is here. He has been one of the active members who has been doing that.

With still that much effort we couldn't include Saraiki language within our scope because we couldn't find relevant experts for that so that is something that we are going to put into the next final version that we're going to submit to integration panel.

Coming to the feedback that we received, we actually had two rounds of feedback from integration panel. First was on 8th September 2017 and the second was on 11th October 2017. The feedback that we received on 8th September we modified the draft and resubmitted it back to integration panel on 11th

October, and then we got an additional feedback from integration panel on that for which have cursory analysis done but in some cases we do not have concrete responses to some of the feedback received from the integration panel. We will do that in due period and submit it to integration panel soon.

So the previous feedback that we received from integration panel, these are the major bullet points. There were some minor changes in wordings. Wherever required, the changes were directly made. Wherever not possible, more elaboration was added taking into consideration what might have triggered the change in the wording.

Then more explanation of certain aspects was demanded by integration panel about the rationale behind inclusion of a certain set of characters. This was to be added. We will do that – major modifications suggested to the XML

In the initial submission when we submitted the document to integration panel, I must admit that we did not use the online tool for XML validation so it had some validation errors and in the recent XML that we submitted we used that extensively and I hope that those many validation errors were not there.

Then harmonization of similar terms with a unique name. Generally, this happens when we are talking within the community. For us a vowel modifier or a Matra or Eyelash Reph

or Eyelash Ra, they actually mean the same thing so we inadvertently happen to use them in an intermixed way but thanks to integration panel they looked at it from a neutral point of view and they gave us a suggestion that you better harmonize the terms, which is a valid suggestion and we incorporated them into that and so far as possible we have harmonized the terms.

In the initial submission we had given a reference to each of the characters which was a standard but it was behind a [pay wall] and so we changed it to a more open accessible standard and in this recent submission we have used Omniglot references mostly.

Then request for consideration, there is this interesting case actually where Neo-Brāhmī initially had proposed a Halanta-ending variant which was of the nature that the last Halanta gets mapped to Hanal so it matches with a same label without the Halanta.

And the integration panel feedback received was if you can, can you reconsider the case because there is no precedence to the LGR tool itself whether it can handle such a case. However, they said that if it is absolutely required let us know. We will see how it will be accommodated. However, when we went back to the community and saw do we really need to propose this as a

variant, the general consensus was that we should not include this so we've taken out that variant.

Coming to the second set of feedback received on the draft on 11th October – this is the draft date actually, we received feedback in the last week itself – there are certain clarifications sought on Sanskrit characters. Sanskrit being a very old language, there is an assumption that some of the characters if required for Sanskrit then they are of archaic use. We are still discussing how to take a call on this. We will be adding more text elaborating the requirement of certain characters. So changes in wording, wherever they are possible, we will make them, and wherever not we will again add more elaboration.

There were substantial comments on 0944 character which is vocalic [inaudible] sign, long [Re], and this actually is a mistake in the code point repertoire. Whereas in one section it says that we're going to disbar this particular [inaudible], it has come into the code point repertoire. Again, we'll discuss this within the community – what exactly is the action that we need to take about this particular character and we will rectify it that way.

And then there were some unaddressed points during the last review process, mostly because we were focusing majorly on getting the document done and sending it across. Some changes were considered, it's just that they were not notified to the

integration panel so the integration panel has resubmitted those comments to us to consider and we will in the next submission which is final. We will address each and every point that integration panel has made to us.

So this Halanta-ending variant was there, which I earlier discussed, which IP suggested for us to reconsider and we ended up removing that. However, IP wants us to document this case, if at all, why this requirement was failed and if the variant is not needed, why it is not needed, so more elaboration needs to be added on this part. We will definitely add that. Then there were some small technical changes required for XML. We will do that.

There is a discussion on some Nukta characters as required by Konkani because they are being shown on the Omniglot. However, this is something that we need to look at because the Konkani community per se does not say that those combinations are required, even if they are mentioned on the Omniglot, and this is something that we need to as a GP also take a call on whether we need to solely rely on Omniglot as a reference or need to do something else. We definitely would like integration panel's guidance on that particular aspect.

Then in the XML, there are descriptive sections which is like the final thing that makes the XML readable by itself. We had not included them yet because again the focus was getting the

document done and XML was to be populated later on so we will add all those descriptive sections in the XML once we are more or less finalized on that.

The same goes with the summarization of the Whole Label Evaluation rules variants in the XML. This will be done in the final submission.

Future plan of action, we are almost done with the Devanagari LGR. There is a good amount of momentum within the generation panel to get the scripts going so we are into that phase. Even in the morning today as Dr. Data was saying, we had a meeting and we have taken a resolution that we will within the next three months at least get some set of scripts going and the other ones soon after that.

The scripts for which we are all set to begin work after the Devanagari are Bengali, Gujarati, Gurmukhi, Kannada, and Telugu, because we already have the experts on board with those. So that is about the future plan of action. With that I come to the end of my presentation. "धन्यव द !" This is Dhanyavaad [returning] Devanagari which means "Thank you for listening patiently to my presentation." Thank you.

SARMAD HUSSEIN   Thank you very much, Akshat. Very well done. And one thing I think I'm just trying to add in your presentation is Akshat mentioned about the next face-to-face meeting in Sri Lanka. He did not mention the date so that is 13th, 14th, 15th of December so if anyone of you is interested to join in, you are more than welcome to block these dates and please attend. Now questions if someone has any.

One more thing actually, a few messages I got that they needed documents so just for the record the document is already available on the schedule on icann.org and with this session if you go to then you'll be able to download that presentation from there. That's available.

UNIDENTIFIED MALE: One just quick remark. When I was talking about the Maithili language, I heard so from those members they use a Mithilakshar. Is that considered or Maithili language is only in the Devanagari so that there are two ways of writing Maithili language.

AJAY DATA: Mithilakshar actually has its own script if I'm not wrong, which is probably not part of the MSR yet so we cannot start working on it. Eventually when the MSR considers that to be mature enough

ICANN 60
ANNUAL GENERAL
ABU DHABI
28 October–3 November 2017

to come into the MSR and if there is enough community momentum, I guess that will come. Mithilakshar actually is the script for Maithili language.

UNIDENTIFIED MALE: If we don't consider right now, we'll have that Maithili in Devanagari ultimately so –

AJAY DATA: This is not given in the scope to –

UNIDENTIFIED MALE: My concern is that look after and then, as Akshat said, so you have gone through that Mithilakshar and it's not in the scope so you left out, so that answers my concern. Thank you.

UNIDENTIFIED FEMALE: I have one question online from Lucas Demora. The questions is, "What is the mount of population that is expected to be privileged with the Devanagari insertions?"

UNIDENTIFIED MALE: That is very large. That is, to be precise, 4.36% population of the world. It has top six… so it's just Hindi so Devanagari is going to be actually much more. I'm just talking about Hindi right now

ICANN 60
ANNUAL GENERAL
ABU DHABI
28 October–3 November 2017

because larger part of this so it is going to be probably more than 6% of the world population. So I was just going through today the Hindi is the top sixth language in the world in terms of population which uses as a user community. So this panel is doing a very significant work which is going to influence and address as far as population is concerned.

AJAY DATA: And Napali is around 0.8 to 0.9 so you can add on like this.

UNIDENTIFIED MALE: So it's all of the south Asian population, almost all of the South Asian population.

[AKSHAT JOSHI]:   So maybe if it is [inaudible] for people and just to answer this, I have gone through the Wikipedia today and taken note of it. Mandarin tops it in terms of the population which is 14.1% and second is Spanish which is 4.85, third is English which is 5.52, and then Hindi stands fourth, that is 4.46, and then Arabic and then Portuguese and then Bengali which is 3.05, which is also part of our panel, and then the Russian. So these are the top eight and out of these, two stand within this panel right now which is 4.46 and 3.05 within the scope which we are doing right now, and maybe another percentage if I go further.

EN

UNIDENTIFIED MALE: This number whatever rates calculated, it's –

[AKSHAT JOSHI]:   This is as per the population of 2010.

UNIDENTIFIED MALE: That population which is taken together who either speaks or understands these new Brāhmī languages.

AJAY DATA: Speaks, the speakers.

UNIDENTIFIED MALE: Only the speakers at the motherland in the mother language.
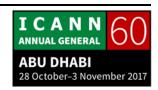
AJAY DATA: Yes.

UNIDENTIFIED MALE: So India has a population of 1.2 billion. Bangladesh has a population of some I think 120 million or so. Sri Lanka and all so all put together it's about 1.5 billion which is speaking that. Now that comes to about –

AJAY DATA: You're right, one billion but one billion is covering 22 official languages not just Hindi.

UNIDENTIFIED MALE: No, I'm saying those covered under –

AJAY DATA: Yes, so NBGP will cover everything of course yes, of course yes. Yes, Anivar.

ANIVAR ARAVIND: Do we have any IBNF tester for giving a string and testing if it's accepted by this or not? Because if we have a tester, that will help for com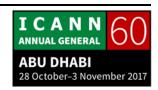mon people to test how good it is and the rules are proven or not, because without a tester it will be very difficult for people to understand this.

[AKSHAT JOSHI]: The XML tool which has been developed by ICANN, once we submit the XML with the relevant rules, we can upload it and then that kind of acts as a tester for you.

ICANN 60
ANNUAL GENERAL
ABU DHABI
28 October–3 November 2017

EN

ANIVAR ARAVIND: I'm not telling about XML testing. Basically it's just for entering a language string alone and getting results of it as accepted string or not.

[AKSHAT JOSHI]:   Yes, so that exactly is the output of that XML validation tool. It comes up with a set of valid labels and what is invalid. It additionally comes back with a response saying why this is invalid so this is already there, yeah.

AJAY DATA: Any other questions? And it would be great if we can have some feedback from IP and if you were to have some comment on this.

UNIDENTIFIED MALE: In general we're impressed. We're impressed by the work you've done so fast on getting all those. Yes, we see that for the Devanagari LGR it's pretty close to be done. We can see that. There's some minor point but I don't see majors so I would expect that to be completed, go for public review pretty soon.

We have been converging on the feedback and you have been accommodating the feedback so we don't really have really a major issue at this point. It's just minor points. Like one of them, for example, was confusing contacts on action rules.

EN

It's also true that to some degree WLE is a bit confusing to some of the actions and some contexts on some things. We could be clear on saying what is a context rule and what is a label rule because they do apply differently and obviously you don't want to put the context in the action section, but I'm sure you'll take care of that.

In other words, we have to rely on you to define what the running system is and then our role is more to make sure that what you write is what you put in the XML, so also consistency between what you say on your documentation and what you put on the XML itself.

We're very pleased and we're waiting for the other one to come I guess and see how they look like and it's going to be challenging a bit for us to do the integration of so many LGRs at once but that's what we're there for. I don't have really much feedback at this point except that we have been working with you and we're pretty happy with the way it's going.

AJAY DATA: Very much relieved to hear your words.

UNIDENTIFIED MALE: My only concern is maybe Devanagari is doing fine without the join or non joiner and I'm kind of curious to see what's going to

happen in some of the other languages or scripts where it's a bit more pragmatic, I guess. I would like to see obviously but at this point like I know for some of them you – because it's part of – for example, if you are doing a second level you'd probably have more freedom on maybe [inaudible] useful from a documentation point of view not probably on this document now but at some point in the future you may want to extend a bit on your document to explain if I could use a zero [inaudible] and non joiner, what are the consequences on the matter.

Because some people will use a root document as kind of it's a definition of what you should do and it's not exactly true because in fact if you allow the join and non joiner, you can do a bit more things especially on the second level. But that's not obviously for this context of what we're doing now, the root zone. You're doing the right thing.

AJAY DATA: Thank you very much, actually. This is like passing an exam. And actually I would like to bring on official record that I'm extremely pleased and we all should be very pleased that our joint secretary representing the government is sitting besides us and taking note of what we are doing.

That is not only [inaudible], sir, that is also very helpful that we are able to update you and you're able to participate and see

what we guys are doing, because if the government supports these kinds of initiatives, then obviously it can turn around things to be much, much matured and much, much better. Thank you very much from the NBGM Panel sir. Thank you very much.

UNIDENTIFIED FEMALE: Could you speak to the mic and please state your name before speaking?
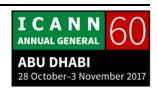
[SHITRA DEV]:   This is [Shitra Dev] from IMAI. Akshat, you mentioned something about having some extra characters. Could you elaborate on that, please, including extra characters beyond the regular number of consonants?

AKSHAT JOSHI: Yes. Those are not extra characters, those are characters which help peculiar requirements in terms of where they join and where they don't. Those requirements are in deviation from the general requirements of the characters that Brāhmī script uses.

Why this is there? There are many reasons. One of them could be the way Unicode has encoded them. There are many factors which affect that but however we should consider that as our

digital given that we have these characters and they need to be tackled separately.

So again when we will be discussing the other set of scripts which have these characters, we will definitely need to be more vigilant about we cater to those characters specifically. And those are very important characters for those particular scripts because that's the reason they were encoded into the Unicode and we cannot ignore them.

ANIVAR ARAVIND: Anivar for the record. Just continuing with what was [known to be] zero with joiner and non joiner so [inaudible] to it. That's one part. Second, I'm not expecting our panel to be incorporating the [buxin] Unicode. There are [inaudible] Unicode which is resulted use to because of a bug in Microsoft font and Microsoft even can't render the ender which is Chilla, Harland and Ra, still on Microsoft Windows 10 so that mistake of Unicode is inherited by a IDNA anyway even though it's not linguistically correct. So I'm not expecting this panel to solve that problem at the same time. What will be a better way? I see that one IBNF rule is totally dedicated to that bug now so what will be a better way for you in this case?

EN

AKSHAT JOSHI: Answering your first question which was about the zero rejoinder non rejoinder requirement for LGR, as [Michele] also said, he's not expecting that to be part of the LGR because this has a very much restrictive scope. However, we definitely need to document that because in some of the scripts they're absolutely needed to represent some of the phonetic constructs and we should document it as such that this being a very restrictive case.

However the second levels should not take this LGR in total and they can if they want to introduce zero rejoinders and non joiners into their LGRs. So that's one.

And actually I could not get why you said that this was because of a Microsoft bug that the Chillu Halanta consonants construct because it could be the case that currently the implementation doesn't support that but that as I see should not deter us from including them as a valid construct.

Because, for example, if you could see in one of the slides of variants my Kashmiri characters were not rendered properly because support to the Kashmiri characters on the majority of the operating system has not yet come. It's because there is a very less demand for it.

ICANN 60
ANNUAL GENERAL
ABU DHABI
28 October–3 November 2017

However, that doesn't mean that we should not as a community support that so can you just elaborate a little bit more about the Microsoft bug thing here.

ANIVAR ARAVIND: History on that case is well documented even within the Unicode mailing list in 2011 by [Sandoshto Tingel] who is also part of the panel and [Vicky Midea], language engineering team engineer. So the histories coming from the Chillu as per [inaudible], but Microsoft [inaudible] phone had a bug which they produced in the wrong format. Instead of the Halanta format there produced it with… It was supposed to be na Halant Ra but they produced letters N Halanta Ra. And instead of fixing the bug, they fixed the send it and they accepted it and [inaudible] so. So this is the historical background in which it is coming and this is being reflected in like the same reflected idea as well so how to find the solution for that is the question.

UNIDENTIFIED MALE: I wanted to comment on that a bit because I don't think – in many cases when there's issues like that the best place is [to rant] after Microsoft. It's more like to get sure that the appropriate behavior rendering of sequence of characters is well-documented and is agreed by, for example, UTC. When you call UTC it's fully aware of these kind of design issues discussed

and to some degree improved because that's where we – now I'm talking a bit on behalf of Unicode Consortium here. When there's an issue like that and there's something conflicting, hoping you know what should be done, the best place to get conversions is really on Unicode Consortium Technical Committee, UTC. That's where this kind of decision gets not necessarily made but at least modified and documented and then the vendors, including Microsoft, can do the right thing because they know that a consensus is reached on some of those complicated issues sometimes.

On conflicting, not everyone agrees. I've seen a lot in the discussion about the Indian script and the UTC and there is a lot of differences on opinion on many of those so I just want to make sure that I don't want to just blame one vendor. It's really more like getting to show that we have a consensus on how to do things because it's not a simple matter.

ANIVAR ARAVIND: The point is how we can find – in languages using [inaudible] in two ways now. Because of Microsoft bug, some people use it in the wrong way. At the same time any syllables analyzer will fail that kind of a role. It's very similar to the IBN of testing but at the same time if one type does register automatically, it creates a

ICANN 60
ANNUAL GENERAL
ABU DHABI
28 October–3 November 2017

duplication issue for adapting so both are conflicting each other because both are usually similar.

AJAY DATA: Can I come back on that? I'm kind of relieved that you say that this is something which is probably outside the scope of this panel. It's not directly about this panel. However, we definitely need to –when we're drafting the LGR we are actually sitting on drafting the IBNF rules.

So we as a community can take a call whether we need to support both the [forms] and harmonize them as a variant or… I'm saying this because I'm not exactly aware of the script. However, there are mechanisms to deal with and we need to take a sense within the community which is now a widely accepted form. Should we just support that one to avoid further confusion or should we debate further to narrow down on what is correct than widely accepted. Or we go in a more pragmatic way where we allow both and harmonize them as variants which I won't advocate because that appears like mitigating the problem for now but it will let the problem remain for a long time. So these are the kinds of issues that we expect probably the Malayalam community to debate vigorously and get the solutions ready for us.

EN

SARMAD HUSSEIN: Can I add here to close this discussion here is that if we can request Anivar to put up a small paragraph of the examples and share with the community and then on another fortnightly call we can discuss this out and take it further, I think.

AJAY DATA: So maybe it is then just good to extract and put this much so that everybody can just read the relevant part of it. That would be nice so that everybody can apply his mind and we can reach a consensus. Any other questions?

So with this, before we close, I would like to take this opportunity to the community that we are going full throttle for another eight scripts and we have a huge commitment in the morning with the entire base to complete them in parallel as fast as possible.

We are not going to take years for sure. In months it is going to be completed and I would like to name those scripts officially here, that is Gujarati, Gurumukhi, Bangali, Oriya, Tamil, Kannada, Telugu, and Malayalam. If you know anybody, if you would like to participate there, if you would like to contribute there, you are welcome with open arms and please happy to join. Message NBGP panel through subscription or to Sarmad, to Pitinan, to me, anybody. We will be happy to take your capabilities and contribution and as we have a general

ICANN 60 ANNUAL GENERAL
ABU DHABI
28 October–3 November 2017

consensus that your name also becomes parts of the LGR. Thank you very much.

[END OF TRANSCRIPTION]