ICANN83 | PF – String Similarity Evaluation in the New gTLD Program-Net Round
Wednesday, June 11, 2025 – 09:00 to 10:15 CEST

PITINAN KOOARMORNPATANA: Hello and welcome, everyone, to the session String Similarity: Evaluation in the New gTLD Program Next Round. My name is Pitinan Kooarmornpatana, and I am a participation manager for this session. Please note that this session is being recorded and is governed by the ICANN Expected Standard of Behavior and ICANN Community Anti-Harassment Policy.

During this session, the questions and comments will only be read aloud. If submitted within the Q&A pod. If you would like to speak during the session, please raise your hand in Zoom. When called upon, virtual participant will be given permission to unmute in Zoom. On-site participant will use a physical microphone to speak. Please state your name for the record and speak at a reasonable pace. I will now hand the floor over to Sarmad Hussain. Sarmad, over to you.

SARMAD HUSSAIN: Thank you, Pitinan, and welcome everyone to the session. We will be talking about String Similarity Evaluation in the New gTLD Program: Next Round. What we'll do is try to cover three different topics. We'll actually take questions during the presentation. So, if you have any questions, just raise your hand in the chat room,

online in Zoom, or please feel free to come up here and speak into the mic.

So, as a starting point, we'll give you a brief overview of the scope of String Similarity evaluation. This is defined through the policy work which has been done by the GNSO. And then based on that, we are actually proposing, there were obviously some challenges with String Similarity to address those challenges, we'd actually proposed a solution. And based on that solution, we'll share progress on-- The solution has two parts, data as well as tooling. So, we'll give you an overview of where we are on data and how we have collected it, as well as where we are on the tool. And as I said, please don't wait till the end of the session for Q&A. Please feel free to jump in whenever you have a question.

So, let's get started. The String Similarity evaluation for the new gTLDS comes through two policy recommendations, or two sets of policy recommendations. The first one is the SubPro, which has a Section 24 on String Similarity. It basically reaffirms the work or the String Similarity evaluation, which was done in the 2012 round, and says that strings must not be confusingly similar to an existing top-level domain or a reserve name.

And also, Affirmation 24.2 importantly confirms or reaffirms that this similarity is bound or restricted to just the visual similarity. So, there can be many other kinds of similarity. Two strings can sound the same, they can mean the same, but those are not in scope. The

only thing in scope is the strings, which are visually identical, and this is coming directly from the policy which we have to implement.

And then there was a follow-up work, follow-up policy work by GNSO which was called IDN Expedited PDP. They had two phases, phase 1 and phase 2. In phase 1, which is also now approved, goes into details on what should be compared. It sets a scope of comparison. We'll talk about that in one of the next slides. As well as it says that the String Similarity Review must be decided by a String Similarity Review Panel. So, it has to be eventually a manual process. It cannot be done automatically through a tool. And also that the panel should have some guidelines to follow. So, there are a few constraints put on the panel as well, just to make sure that the results are predictable and transparent.

So, this slide explains the scope of String Similarity review or evaluation, which needs to be done. On the top there, on the right, is the applied-for string. And one of the new, I guess, more recent changes in policy is that it now allows for not just the strings, but variants of strings. Variants are calculated using the Root Zone Label Generation Rules, which have been developed, published through support of the community. And variants are defined as those which are generated by the Root Zone LGR.

Variants are then subdivided into two kinds of variants. Allocatable variants, which are those variants which an applicant can apply for. And then there are blocked variants, which an applicant cannot apply for, but they are reserved and cannot be allocated to

anybody else either. For a string, the allocatable and blocked variants are also defined by the Root Zone LGR. They are not determined by the applicant. Applicant can only choose from the allocatable variants, which ones the applicant may want to apply for. But in any case, so we have the primary gTLD string, which is the string being applied for. It has allocatable variants and it has blocked variants.

Similarly, so as per the policy, all these strings would need to be compared with the existing gTLDs, existing ccTLDs, also block names, and also gTLDs, which are in process from the previous round and any ccTLDs which are in progress of evaluation. So, those are the different-- And not only the strings, their strings, but also their allocatable and block variants, as calculated through the Root Zone LGR.

Suppose there are 100 strings applied for, then there are, suppose, 1000 TLD strings and reserve names. The comparison is no longer between 100 and 1000, but it is between 100 strings applied for and their variants, versus maybe the 1000 TLDS and block names, let's say, and their variants. So, the number of comparisons is extremely large. And that was one of the challenges which we had to deal with for String Similarity this time around. In the previous round in 2012, the comparison was just the strings which were applied for and the strings which were actually delegated, for example.

So, the problem space has significantly increased. As you can see, the primary gets applied for. If you look at the first column under

primary gTLD string, it has to be compared with the primary of the comparison categories, but also their allocatable variants and their blocked variants, and so on. So, all of these things are in scope. The only thing out of scope is that blocked variants versus blocked variants. So, that's the only thing which is not in scope.

So, what we have done is we've started developing guidelines on what could cause confusability. We had an initial draft of this, which was published for public comment some time back. And in that draft, we basically identified some of the categories, for example, which could cause String Similarity issues or challenges, for example, uppercase versus lowercase underlining, so we've discussed all the very different kinds of challenges.

And one of the things which we identified in that particular public comment was that given the larger scope of comparisons now, even though eventually the comparison or determination has to be manual, for String Similarity evaluation, it may be a good idea to have some kind of tool to support that manual process. So, if two things are completely distinct, so, for example, if you're comparing a Chinese string to an Arabic string, there is a low likelihood that those two strings are going to be confusing. And therefore, maybe we do not want to give every pair of strings and equal weightage as far as comparisons are concerns. Because there are thousands and thousands of these comparisons.

So maybe what we want to do is we want to bring up some of those strings or pairs of strings which are possibly more contentious. And

then eventually, the panel can review them and find out what are actually the contention sets. So, what we did was that, or what we suggested in the initial guidelines we published, was that what we'll do is we'll get experts from all the different scripts which are supported, 27 scripts now in the Root Zone LGR, and ask them to identify all the confusing characters within the scripts and across scripts. And then we use a tool to use that data and then the string information to see whether it creates any possible confusions and highlight those for the panel to eventually manually review and decide. And we'll share those details with you here, the process.

So, before the tool could work, you would need to have data, and that data would need to come from experts. So, what we did was we actually reached out to 27 different experts for 27 different scripts which are in Root Zone LGR and requested them to look at the script and highlight whether any of these factors are relevant for any of the characters which are in those scripts.

So, we wanted them to compare all the characters in their script with all ASCII, of course. Also, compare them with all other elements or characters, for example, within the same script. Also, compare these across the script. So, for example, there may be some similar looking glyphs between Greek and Latin script or Armenian script and Cyrillic Script and so on. Maybe we don't need to compare them with all scripts, but at least those scripts which are in the same script family.

And then we also ask them to look at other challenges like upper-case, lower-case forms. In some scripts in complex scripts, they have conjunct forms, where two characters when they join, they form a third shape, which is different from the individual two characters. And so, we want to look at those kinds of things as well. And then there are simple shapes which occur in many scripts like a round O or a simple line like I or L. So, we want to identify those simple things which are prevalent across multiple scripts and underlining. So, those were some of the things we actually asked our experts to look at when they were identifying.

And we requested them to categorize any similarity in actually five levels. Level 1 is when they think-- Well, all the variants which were already identified are categorized automatically as Level 1, because they, by definition, are the same for the scripts. But we define Level 1 as identical or near identical. And there were maybe a few more which were identified, but largely this is variants which are already identified by the community.

Level 2 was highly confusable, and we put in some examples here on the slides as well for you to see. Highly confusable are strongly similar and near homograph, but not quite. Level 3 basically are similar, not strongly similar, but they still have a high probability of being perceived similar by people from that script community, or one of the two script communities. Versus Level 4 are weakly similar. They may be confused by some people, but not a majority

of people using that script. And Level 5 was those which are just totally distinct.

So, those are the five categories we asked the experts to put their data in. So, each of the experts went through all the repertoire, all the characters in their script, and identified them, any pairs which should be either Level 1, Level 2, Level 3. So, identified the pair and also marked whether they're Level 1 or Level 2, or Level 3, and so on. And gave the reason as well. So, there's a rationale associated with their categorization. So, that's the data which we've collected.

This is the status of the data. So, we've actually gone through most of the work. We are finalizing internal review and we're ready to publish for public comment. So, you will actually be seeing the public comment coming out very soon. And I guess one of the things we wanted to highlight was to sort of see eventually since this is work which is done by just the expert of a script, we really want the larger community to look at it and comment on whether they agree that two characters, for example, which have been identified as similar, are indeed similar, and whether the categorization of that similarity on the scale of 1 to 5 is something you agree with or not agree with.

We will take that input, of course from the community, from everybody, through public comment, and then finalize the data, which then is going to be used into the similarity review process. And this is the list of scripts which we are covering. Again, it's the

same list which is in the Root Zone LGR, because eventually only scripts which are in the Root Zone LGR can be applied for.

These are names of the experts who actually have worked on the different scripts. Obviously, we'll publish that along with the data as well. Most of these have been actually involved in the Root Zone LGR development work as well for these scripts. But again, at the end of the day, we will take it through a public comment process to see if there are any changes. We would want to make sure that those changes are incorporated in consultation with the experts.

Okay. So, what we are now planning to do is we are planning to publish this data on String Similarity which we've gathered for public comment. We are going to be publishing the data in two versions. The HTML version, which is the human readable version, and then the corresponding XML version, which is the machine-readable version, which we'll be using. The HTML version is created automatically from the XML, so the two should be aligned, but we will actually be using the XML in the tool. HTML is, of course, there for you to review and provide feedback easily. And it includes these confusable sets and many more details, but I think the main focus which you want to do and review is whether the sets which are identified as confusables, you agree with them, and if not, you should let us know.

So here is an excerpt of one of these from one of these tables, it basically lists the Cyrillic small letter SHCHA, I can't pronounce that properly, Armenian small letter PEH, and Ethiopic syllable SZAA in

the same set, for example. And for each pair, it actually lists whether its similarity is level 1, 2, 3 or 4, for example. And as you can see, it also provides, in the final column, a rationale of why it is being done.

An interesting thing which actually happened was that when experts provided this data, they could say that A is similar to B, and B is similar to C, and we get that data, but they don't say anything about A and C. So, what we've done is programmatically done a transitive closure and included, for example, in this case, A and C in the list as well, and marked it imposed. That's like the middle, if you see the comment in the middle, it says imposed. And we've given it Category 4 by default, because it was not identified as similar by the expert.

But that's a default category, and that's the special, because that's introduced mechanically. Especially the ones we will actually be asking the community to pay more attention to. Because they were not identified by experts, so we consider them not similar. Because if they were similar, the experts should have identified it. But since these are mechanically generated, we want to make call special attention to it. So, that's one thing which has been mechanically added into the data, the transitive closure, and we marked that explicitly for everybody to review as well.

For each of the pairs, we are giving rationale for why each of those pairs are considered similar. Not us, but the expert is giving that rationale. And these are some of the reasons of those rationales

that it could be-- There are multiple reasons, which are I guess explained here. We don't need to go through it. But when you get these documents for public comment, they are actually documented.

I guess this is what I was talking about as well. We have this imposed category which we've introduced, this was mechanically added. And this is just to make sure that the transitive, all the similarity sets are transitive. And if there are things which are missing, they are added, but with Level 4, which means--

By the way, I probably didn't explain that when we are going to compare strings, what we are suggesting is levels 1, 2 and 3 will be considered similar, and level 4 and 5 will not be considered similar. So, by default, the imposed categories are given Level 4 because they're possibly similar. But they're not quite because they've not been identified by experts. So, they'll not make the cut unless we get input from either experts or from the community that some of them may need to be changed from Category 4 to categories 1, 2, 3. So, that sort of lays the basis.

As far as the feedback request for the public comment is concerned, we will request you to identify if any similarity pairs are missed from this list for your script. If so, identify that pair and the similarity level, and the reason why it should be that similarity level from you. For pairs which are identified by experts already, review the degree of similarity marked and either confirm the value or provide the updated version with the reason. Any imposed

mappings, these are the ones which are mechanically introduced for transitive closure. If they receive a category of 4 by default, we will request you to review and see if this should be updated to a value between 1 and 3. And, of course, any other feedback is very welcome, but those are specific inputs for the data.

As far as the next steps are concerned. We are now in the final stages of getting the data together. And we've already gone through and done a couple of rounds of review with experts. For us, the next step is to open this for public comment. And that should be coming within the next few weeks. And then once we get input from all of you on public comment, through public comment, we will take that input back to the relevant experts for those scripts, discuss that and finalize the data based on the input we have received. And then we will actually publish the data as well and then integrate that into the tool. So, you would actually know what the final data is from String Similarity perspective.

So, that brings us to end of the first set. And I want to maybe break here to see if any of you have any questions so far before we get into details of the tool. Yes, please. And there are mics up here. If you want to come up and ask a question, you can do that. Or you can post your question in the chat. Yes, Werner, please.

WERNER STAUB

Werner Staub from the CORE Association. I have two questions. One of them is, are similarities that arise from combined letters, such as RN with M? Are those part of the scope?

SARMAD HUSSAIN

Yeah. So, when we requested the experts, we didn't ask them. What we said was that they have to compare elements in a script. Element could be a character or a combination of character. So RN, for example, is identified as something which is similar to M. So, yes, you can actually have code points or code point sequences, similar to other code points or code point sequences.

WERNER STAUB

Okay. My second question is about the definition of similarity. I actually usually say there's no such thing as visual similarity to a trained user. Everybody's a trained user in writing, in their respective script. Because we learn, as we learn to write, we learn that things that look similar are different, and the things that look different are identical. That is the process of learning. And, of course, in the similarity context, the first experience in ICANN was Chinese script, where there were really different characters, not visually similar at all, that are defined as being the same.

So now we are used to that in the European languages because of uppercase, lowercase, they're really, really different. But we consider them to be equivalent. And, of course, in the online world, we've all been trained with new similarities that they did not

consider before, which was that everybody leaves out the accents, we just leave them out because that's the only way. That's almost like the definition for every user in practical life for the last 20 years, you just strip off the accents if you want to do something online. And so, do you consider this learned similarity as part of the scope?

SARMAD HUSSAIN

Yes. So, we published these guidelines, which I was referring to I think, almost a year ago. And one of the things we discussed in those guidelines is how do we, precisely the question, how do we define similarity? And who's the, I guess, beholder there? And what we said was that eventually, when we're looking at similarity, we're looking at it from the perspective of somebody who's familiar with the script. Not an expert of the script, but familiar with the script. For example, I don't read Chinese, so we don't expect similarity from a criteria of when I am trying to read Chinese, but from a person who reads Chinese scripts is reading Chinese. And then if those characters are considered "probably similar", then they're marked as such.

And some of these details are actually discussed in those guidelines as well. But our definition of what is similar is precisely that, that a native reader reading it not casually, but reasonably. Because sometimes when you just skim over something, you can find something confusing as well. But if you really sit down and read it, you can differentiate. So, our definition was that it has to be a native speaker, and they need to be reading it reasonably carefully.

They're not just skimming it. And if at that time as well, they find something which is probably confusing, that is what should then be marked.

And this is marked not by us, but obviously by the Native script expert who we've actually hired and who actually worked on Root Zone LGR as well, defining variants and all these other things. So, they're very familiar with this space. So, that's where the data is coming from. But again, it is going to be open for public comment, so please do take a look at it. And if you don't agree with something, please give us feedback. Any other questions?

PITINAN KOOARMORNPATANA Well, we have one hand in the room.

SARMAD HUSSAIN Yes, please, Pitinan.

PITINAN KOOARMORNPATANA Satish?

SATISH BABU Hi. Thanks for the presentation. This is Satish from At-Large part of the PDP currently sitting in India. Hi to everybody. I'm kind of trying to raise an issue from the end user perspective. Many of the end users are not experts. They probably are general users. They can read, they can understand, but they may not be able to look at

the final aspects of the glyphs. So, in the public comment, are you recommending any particular process for non-experts on how they are supposed to look at this task of identifying what is similar and what's not similar? That would be very helpful for them because by and large, the vast majority of At-Large are non-experts. Thank you.

SARMAD HUSSAIN Thank you, Satish. That's a good suggestion. Maybe I think if we can follow up and see if we can work with At-Large and other parts of the community to see if we can get this public comment, I guess, shared more broadly with the communities, I think that will help. And so, what we can do is we can follow up with you and maybe ALAC team when this comes out for public comment to see if we can also channel this to broader local community communities, actually globally, so we'll follow up. Thank you for the suggestion.

SATISH BABU Thanks.

SARMAD HUSSAIN Any other questions? Okay. I don't see any. So, then let's move on to some of the details of what the tool does. And I'll request Pitinan to come in for this part and present. So, over to you, Pitinan, and let me know when you'd like to move the slides forward.

PITINAN KOOARMORNPATANA Thank you, Sarmad. Hi, everybody. It is Pitinan here. So, I would share some detail on the tool and also go through some of the examples, given we have some idea of the data now, so we go through that in this section. First of all, where these two will be fit in into the process. So basically, in this diagram explaining that, TAMS is shortened from the TLD Application Management System. So basically, this is the main system that will be used during the Next Round.

So once the round is closed and we have the set of apply-for strings ready, those set of apply-for string will be sent to this Settings Similarity evaluation tool when it comes to that process. So, in this tool, they will have to pre-install the data, so SSE data. This is what we are gathering from the experts. And then there's pre-existing lists, existing TLDs, existing ccTLDs, block names, reserve name, and so on. So, this will be pre-installed in the tool.

Then it will take in the apply-for string as an input. It will identify the possible confusable set based on the data that we have, and then it will generate the system report, which will include all the apply-for string. This report, we will share some examples how this look like later, this report will be shared with the panelists and String Similarity Evaluation Panel. They will take a look and maybe, if they want to see more detail, for example, in the tool, if you have some default threshold of what is the similarity it will generate.

If the panelists think maybe it should change some threshold and see is there anything else to become in the list or left out of the list to support their decision, they can go back to the tool and generate the custom report based on what they would like to see. And once they have sufficient data to make decision, it will finally go to the panel managers, who may need to gather the input from the panelists and manually update the report to be the final report. And this final report from the panels will be feedback to the TAMs to move on to the next steps. So, this is how the tool will fit in. The tool will not make the final decision. The key is this is to prepare the possible confusable set and present to the panel, so the panel have information to work on and make decision.

Next, please. Now the tool. What we plan is to make this tool available to the panelist. It will be available to website. The panelist will need to have again account to authorize in and log into the tool. The tool will incorporate all the pre-loaded data and then it will compare the apply-for strings with all this list, the all other apply-for strings, existing gTLD, existing ccTLDs, the requested IDN ccTLDs that is still in process, the previous round pending gTLD, the list of two-character ASCIIs, and also the block names. And then the tool will be able to detect the potentially confusable levels. Within those confusable sets, it will calculate the visual distance between the members in the set. And lastly, it will also list the score and show the ones within the threshold to see in the report for the manual review.

Next, please. This is the example of the structure of the output. So basically, when it's fired at Label 1 and Label 2, for examples, potentially be confusable, it will be put into the same set. Within the set, each pair of the members in the set will be list in one line. It will go through what is the set and the ID of it. Label 1, the application ID of Label 1, Label 2 and the application IDs if it's the apply-for string or the source, if it's the existing TLDs.

And then, for this time, it's not only compared between the primary string versus primary string. Sometimes the primary string itself is not similar, but through the variance of it, make this two become similar. So, it will also list what is the varying label of this Label 1 and Label 2 that make these two come into the similar set. It will also show the categories. If everything is mapped between the two labels, as the category three or less, then this will be marked as yes. It will also list the score and then some rationale that the tool can detect. The tool will also provide blank columns for be filled in by the reviewer, which is the name of the panelist who eventually make decision, and some notes available there for panelists to put on the final supporting information of their decision.

Next please. As per the timeline, right now we are finalizing the data, and also in the parallel, we also doing some UAT, the User Acceptance Test of the tool development. And now we are finalizing it. And after the fixings and everything, we plan to have the tool ready by September this year.

Next, please. Then moving on some examples. So, this is the first example. In this case, the apply-for strings are similar to the existing TLDs. So, if you see on the left, the list of apply-for string right now, for example, two labels of it. One is B with the hook, a with accent and T submitted by application one. Second is the IDN labels in Thai, two characters submitted by application two. This will be submitted to the tool and then the tool, they have the pre-loaded list. So, in this case, in this example, the B-A-T in ASCII. Example that this is an existing gTLDS and then other labels under ccTLDs that is similar characters, but in Lao.

So, with this information, then the tool will take this, do the comparison with the existing list based on the SSE data and eventually generate the report. In this example, the tool will generate two sets. One set is the B with a hook, -A-T in bat, come into the same set, and the second is the Thai and Lao scripts come into the same set. Just to note that this is very similar to the output of the tool, but it's not exactly the same. For example, in the left part of the tables, because we want it to keep the ability to sort and filter, so the merge cells actually is divided into three rows with the same data in the actual report. But for this presentation, we just merge it so it looks easier to read.

So, let's zoom in for each one. Let's go to the next slide. Let's go to the first set first. The B with a hook in the bat. So, basically, these two primary labels, Column 1 and Column 2. So, Column 1 is from the apply-for string, and then Column 2 in this example is the existing gTLDS. So, in the report, at the column in the Label 1

section, it will always be the apply-for string, and at the Label 2, it will identify that this is from the source of gTLD list.

Next, please. Then behind the scene, the tool will actually try to compare not only the primary string, but the variants based on this matrix in the policy. So, the tool will then generate the variant members of each string. So, on the top, the apply-for string with B with the hook. That one has one block variant label. So, it's two members in the set, the primary and the block variants. Another label is the BAT-ASCII, which is the existing gTLDs. That one also, the two will generate the variant set of it. And for this case, it has one primary, bat itself, and then one block variant, a with acute. So, these four labels two by two will be compared with each other and come into the second half of the report table.

Next, please. So, between these two primaries, then they have four possibilities of comparisons. Just to note that because of by the policy, the block variant and blocked variant shouldn't have to be compared. So basically, the last one, the block versus block, will not show in the report. But the rest will be shown. And then for each one, it will go through a list of code points, and then they map between the code points based on the category of similarities from the data. And then it maps the numbers in the rationales.

Just for example, in the first row, basically the B with the hook, 0253 and ASCII B0062 are marked as the Category 2 in the data. So, in the last column, the first one is 2. And then the next pair, a with a Q to the left and A with the accent to the right, those also mark as 2.

So, the second one is 2. And then the last one is the same T, 00740074. So, that is 1, because it's Category 1, the same thing. So, the rationale is-- 2-2-1. And for all this, the distance, the similarity score has been calculated. This is by the mean square, and the score is there. And then, in this case, everything is 3 or less, so the level 3 similarity is also marked as true. And this process has been calculated for all the pairs.

For the third row, for example, the second code point, the A with a hook above and the ASCII A is marked as Category 4 in the data. So, for the third line there, that rational is 2-4-1, and the Level 3 similarity marked as false. So, all this is the data that will be presented to the panelists. The panelists can choose to adjust the score, for example, like if the panelists submit another request to the tool and say, okay, maybe set the threshold to be less than 0.1, then there will be just two rows shown, for example. So, this is one of the samples of the two.

Next, please. And then, eventually, the panel might pick the closest score and make a decision. That's because of this. Then the primary B with the hook bat and bat ASCII are similar. But again, this is really up to the decision of the panelists.

Next, please. Another string in this first example, which is the string in Thai and Lao. It's in the table. If you can see, sometimes these two characters can be seen as a different font of each other. The feature is very similar. One is a little bit more curvy than the other. In this case, by this example, we just say the Lao script in the label

2 come from the ccTLD list. So, the tool will be able to detect the possible confusible set, and then it will mark that this has come from the ccTLDs.

Next, please. Again, before moving on to all the comparison, the tool will generate the variant set based on the Root Zone LGR. In this case, this tool label doesn't have any variant. So, it will just compare by itself the binary labels. Next, please. So, it's just come to one pair, and the same process has been followed, the rational mappings by the categories, calculating the score, and then just mark the similarity level true-false. And this will be presented to the panelists for their decision.

Next, please. So, that's the first one. Let me move on to the second one. So previously, there was some similarity with the existing gTLD or ccTLDs. This second one is what if the apply-for string is similar to each other. So, in this case, like one of the questions earlier, in the data for Latin script, it's also marked that M and RN are actually similar. So, in this case, the T-A-M, TAM being applied for by application 3 and TARM four letters, being applied for by application 4. This has been submitted to the tool. The tool crunches the data based on the SSE data and generates this report. So, it will be identified that these two should be in the same potential set because of the data that we have from the Latin script expert.

Let's zoom in. Please go to the next one, please. So, for this one on the left part of the table, it will list that tam and tarn are potentially

in the same set. They list the application ID of each of them, and then the source is applied for stream in this case.

Next, please. The tool behind the scene will also do the same, generate the variant of each levels and then do the comparison. So, I didn't put that anymore. But in this case, the three collectors, tam has two variants, and then the Tarn doesn't have any variants. So, for this two by one, you have two rows, and they will be comparing like this as T and T, A and A, and then the M with RN, and the rationale is 1-1-2, as shown here.

And next, please, and then the last example. In this case what if the apply for string not found similar to anything at all, not similar with the existing TLDS block names, and also not similar among each other. So, all the strings will be passed to the tool. And then, if it doesn't find any similarity potential set, then it will just generate the empty report. So, I guess that's all the three examples I have. And happy to take any question, comment. Thank you.

SARMAD HUSSAIN Thank you, Pitinan. We'll open the floor for any questions. Thank you,

WERNER STAUB I'm going to stop again. I have a question about letters that have glyphs that are used in print, but not necessarily in that shape in the URL bar. And there are a couple of examples in the Cyrillic script. For instance, Cyrillic T has a glyph that is exactly, actually

identical to what we will regard as an M. Cyrillic P will be exactly what we regard in Latin script as an N. So, are those part of the scope or is it separate?

SARMAD HUSSAIN

So, one of the factors which we requested experts to look at is also handwriting, for example. So, if two characters are in written form outside the scope, I guess how they're visible to a particular font, but also, if they're written in some different ways, example, of course. And so, there is a visual identical form which exists maybe somewhere else that could actually be considered. And again, I think eventually that is the call of the expert to see whether those two characters will then be considered same and not so. And then there is also another category for them to decide whether there are any other pairs, and they note the rationale.

So, the short answer is yes, but eventually it is going to be based on the expert's call. But it has to be strictly, the rationale has to be a visual rationale. And so, if that was an issue which the expert considered is relevant, they could potentially include that in the data which they provide to us.

WERNER STAUB

So specifically for these letters, the well-known ones are M, N, the U equivalent is the I, is actually a U shape. So, those do not seem to be in the current set of--

SARMAD HUSSAIN I don't know, but I can take those letters and we can check the data and give that feedback. And obviously, I think, as I said, this is coming out for public comment and if you see there's something missing, please do let us know and we will go back and make sure that that's addressed by the experts.

WERNER STAUB And the same question, which is, is there going to be an online tool that we can use to test this kind of equivalent?

SARMAD HUSSAIN So, this was a question last time as well, when we presented at ICANN82 whether the Strings Similarity Tool, which the panelists will be using, will we be able to make that available to the community. And we are actually discussing with our technical team internally. The challenge is that the team's already a little stretched to just make sure that everything is in place for the Next Round in time. If we find the bandwidth that we can make the tool available online for community as well we will do it, but it really depends on the technical team to tell us whether they have that bandwidth.

At this time, we are not planning for it. Our goal is to at least make it available for the panelists and get everything else in order as well, of course. So, at least that's where we are at this time. If we do get bandwidth, as I said, we would certainly want to try to make this

tool available for the community as well. The idea actually is not to make the tool available online, but we can release the source and people can then install it and run it. We are looking into those options, but we cannot promise. We just have to see the schedule of the technical team. Thank you.

PITINAN KOOARMORNPATANA Sarmad, we have Satish in the room. Satish, please go ahead.

SATISH BABU Thanks, Pitinan and thanks, Sarmad. Congratulations on a whole lot of work that has been done, which is good. But I have two observations here. This is Satish for the record. The first is that this is a fairly complicated or a complex subject, which, given the fact that it deals with multiple scripts, not many people would be able to provide an opinion on some of these things.

The second point is that there is a bit of subjectivity here. Earlier we had a tool and the tool kind of gave us some output, which is fine. Here we have a tool and then we have a separate process, which is entirely subjective. The optics of this becomes a bit difficult because we cannot challenge the tool. Of course, we can validate offline, but what comes after the tool, the subjective opinion of this panel of experts, there is no way we can challenge or anybody, I mean, any African can challenge.

And there is therefore, in my perspective, I may be wrong in this, there is a certain loss of optics. The transparency that we expect

from a process like this, especially if there's a contention, we do have-- I'm not sure whether there's any simpler way of handling this, but we seem to have a problem there. It is not of our making. It is the nature of the debate is. But I would like to know, again, speaking from an end user perspective, is there anything we can do to simplify this whole process? Thank you.

SARMAD HUSSAIN

So, I'll answer your question in three parts. So as far as the challenge is concerned, the first layer of challenge is the data which we are publishing for public comment. It's in your hands, so please take a look at the data and tell us if you do not agree with the categorization. We've done a categorization in five levels, not a binary categorization on whether it's similar or not, but five levels, which actually brings in a level of detail. And for each of those, there is actually a rationale provided as well. So please look at data, because eventually the baseline pre-evaluation output which comes from the tool is obviously as good as the data is. So, that part is in everybody's hands. So please provide us that input.

Once the data and pre-evaluation is done, the second layer is actually basically the panel is also not given a free hand. The Policy IDN EPDP Phase 1 says that there should be guidelines developed for String Similarity evaluation, and then panel really needs to work within those guidelines. So, we've actually already shared an initial draft a year ago and based on all the data and work we've done, we are going to come back and redraft and finalize those guidelines,

which are also going to come back to you for public comment. So please take a look at that as well and make sure that there are reasonable checks or constraints or guardrails there for which you would like the panel to follow.

And then, finally, the third part is that the panel decision as the part of the process of application for new gTLDS in the Next Round, there are objection processes and challenge processes. And Elisa volunteered that she'll talk a little more about it. So maybe I can hand the floor to her to talk about what is available in the process from a perspective of String Similarity evaluation and challenges and objections.

ELISA BUSETTO Thank you. Thanks, Sarmad. This is Elisa. First of all, if a string is found to be confusingly similar by the string Confusion Evaluation Panel, that decision can be challenged by the applicant, so there is a challenge process in place already at that stage. But then there will also be a chance to appeal to the decision of the panel for all parties withstanding, which include the number of parts that are listed in the Applicant Guidebook. And I can also send a reference to the exact section of the Guidebook where we talk about this.

If they think that a string that passed the String Similarity Evaluation Review and was not found to be in contention or confusingly similar to the string that Sarmad already mentioned. So, they will be able to challenge this and say that that string is indeed confusingly similar to an existing string or a block name.

And the panel will evaluate that. And should it be found that the string is indeed confusingly similar, the string will be put in a contention set or will not proceed. The string confusion objection, however, goes beyond visual similarity, which is what Sarmad already talked about, but it also includes aural and similarity in meaning, so it expands the scope of the String Similarity Evaluation. Thanks.

SARMAD HUSSAIN Satish, does that address your question? You have a follow up?

SATISH BABU In part, but I think we can take it offline. Thanks.

SARMAD HUSSAIN Thank you. Yeah, just to conclude, At the end of the day with all these constraints and so on, the policy does require that the panel make a manual decision on it. So, at the end of the day, there is obviously some level of discretion which is involved by those experts. Any other questions? We'll give it a minute for everybody to see if they have any questions. Okay. It doesn't seem like it, so then thank you all very much for attending. And we'll close the session now and we can stop the recording.

[END OF TRANSCRIPTION]