ICANN82 | CF – String Similarity Evaluation and New gTLD Program: Next Round
Wednesday, March 12, 2025 – 09:00 to 10:00 PST

PITINAN KOOARMORNPATANA: Hello and welcome to the String Similarity Evaluation in the New gTLD Program Next Round session. My name is Pitinan Kooarmornpatana, and I am a participant manager for this session. Please note that this session is being recorded and is governed by the ICANN Expected Standards of Behavior and ICANN Community Anti-Harassment Policy. During this session, question or comments will only be read aloud if submitted within the Q&A podcast. Questions and comments placed in the chat will be considered part of the chat and may not read aloud on the microphone. Interpretation for this session will include English, French, Spanish. If you would like to speak during this session, please raise your hands in Zoom. When called upon, virtual participants will be given permission to unmute in Zoom. On-site participants will use a physical microphone to speak. Please state your name for the record, the language you will speak if speaking a language other than English, and speak at a reasonable pace. I will now hand the floor over to Sarmad Hussain. Sarmad, the floor is yours.

SARMAD HUSSAIN: Thank you, Pitinan, and welcome, everyone, to this session. Good morning, good afternoon, good evening, wherever you are. So, just to give you an overview of what we will be talking about, initially, I

will present just a very brief overview of string similarity and the scope of the similarity evaluation. Then we'll request Michael Bauland to help us present the effort we are currently undertaking to collect the string similarity data from script experts. This will be followed currently by Julien Bernard and David from the ICANN team. They'll talk about the tool which is being developed to use the data to assist the manual evaluation. And finally the timeline which we foresee. We'll have hopefully time for questions and answers at the end, but please don't wait for the Q&A part. If you have any questions at any point, please either raise your hand in the Zoom or put your question in the Q&A pod or in the chat. Or in the room, just you can raise your hand as well. So with that, let's get started. Just to give you an overview, in the next gTLD round, they're actually going to be multiple checks on the strings. One of the checks is to make sure that the applied-for strings or they've applied-for variant are not similar to strings which already exist or may exist in the root zone. And obviously the motivation behind that is to make sure that the end users do not get confused based on such similarities. The scope of this similarity is purely visual, so similarity can obviously be in hearing, in meaning. But this particular string similarity evaluation based on the policy is very strictly limited to visual string similarity. And there are initially a couple of affirmations in the SubPro which ask that string similarity should be done in between for the applied-for strings as they were for strings done applied for in the 2012 gTLD round. In addition, there are much more details added to the scope and how the string similarity should be done in the phase one final report of the IDN

EPDP which was a subsequent policy process from SubPro. And we obviously encourage you to go and read the specific recommendations there in some way summarized on the screen here, but we would really encourage you to read the exact language of those recommendations in the policy. So just to give you an overview, first of all, I think before we get into the specific details on what is on the screen, it is useful to appreciate the difference between what are called variants and what are called similar strings. Variants are defined, as per the root zone LGR procedure, as those strings which are considered "the same" by the specific communities which are using those scripts. So it has a definition which requires the two strings to be indistinguishable by the community and which would include, for example, homoglyph strings or other strings which are considered similar not only visually, but also in some other ways as well. As far as variant definition is concerned, that is strictly done through the root zone label generation rules as per the policy. And then beyond what is already defined as variants, there could also be additional strings which could be similar if not the same, and therefore they are not defined as variants or identified as variants, but they could still be confusable in case both are delegated for the end users. So in the scope of string similarity evaluation, we're really looking at those sets of strings which are not variants but still confusable. So that's sort of a scope of what string similarity evaluation is about. So based on the recommendation which is in the IDN EPDP, this table identifies the scope of comparison. So on the top, if you see, is the string which is being applied for. It could be the primary gTLD string

which somebody is applying for. That primary gTLD string, whether it's ASCII or IDN, may have actually allocatable variant strings and also blocked variant strings. And as you can see, all those three sets (the primary, the allocatable variants of the primary and the block variants of the primary) are in scope of the comparisons for string similarity evaluation. So the question is, what will these strings be compared against? So if you then look on the left hand side, it lists all the categories of strings against which these strings will be compared against to determine visual similarity. This includes the existing gTLDs, the gTLDs which are applied for in previous rounds but still in process, existing ccTLDs, requested IDN ccTLDs and other applied-for gTLDs in the current round, the block names (and there is a couple of subcategories of block names which are applicable here) and then any two-character ASCII strings which are of course reserved for ASCII ccTLDs. So all those strings will be compared against, and not only those as primary strings, but also all their allocatable variants and all their block variants. So there's going to be a large set of comparisons between applied-for strings and all these other strings and all their variants. The only comparison which is out of scope is that black box at the bottom right, which is the comparison between the blocked variants of applied-for strings and blocked variants of all these other categories which the comparison is being done [to]. You also see these grayed-out boxes. These are boxes or these are comparisons which the String Similarity Evaluation Panel may choose to override or in some ways not do. If the relevant scripts are just so totally different, then there is no chance of string confusion. So in

any case, everything is in scope except blocked versus blocked, and that's based on the policy which has been defined. So we actually published an initial guideline for string similarity review or evaluation for public comment. This was some months ago, but that was an initial version. We, based on additional work and public comment feedback, will be bringing out a new version of these guidelines again for public comment at a later stage later this year. So please do look out for that. These guidelines are for the panel to guide the panel on how to do string similarity evaluation. As you can probably see from the last sheet, there are now going to be many more comparisons because it's not just the strings compared against each other or existing gTLDs and ccTLDs, but also their variant levels and variant levels of what these are being compared against. So there's many, many more comparisons. So to assist the string Similarity Evaluation Panel to manage those comparisons, we are also developing, so to speak, a prescreening tool. What this tool is going to do is use some data which is currently being developed by experts (and we'll talk about that in the next part) and then use this data to create some initial estimates on what could be or could not be similar, and so therefore prescreen some comparisons. And then that data or that prescreening will eventually be used by the String Similarity Evaluation Panel to do the manual check and make the final determination. The prescreening tool is just an aid for the panel, and the panel of course is free to override any prescreening recommendations provided by the tool. So it's just an assistive tool which the panel will use. It doesn't really determine anything. Given that content,

we'll obviously spend some time today on also going over what the tool presents. So with that background, let's see if there are any questions. Otherwise we move on to discussing the data and then the tool part of string similarity evaluation.

PITINAN KOOARMORNPATANA We have [inaudible] hand.

UNIDENTIFIED MALE Thank you, Pitinan. Thank you, Sarmad, for a very clear definition of the string similarity evaluation tool. As you have said, the comparison, since we have included the variant in this, is going to be very large. So do we have an IT tool for this comparison, or, manually, the applicant has to compare for this? Thank you.

SARMAD HUSSAIN As I was sharing, we are actually developing this string similarity evaluation tool which will do the prescreening. So it is an automated tool which helps the panel to shortlist what are possible confusable strings, just because the scope is just so large. But again it's just an assistive tool. It doesn't really make the decision. The panel eventually makes the call on what is similar or not.

UNIDENTIFIED MALE Yeah—

SARMAD HUSSAIN Sorry, we have Crystal Ondo, please.

CRYSTAL ONDO Yep, thank you. I was just wondering, does this have any bearing on plurals versus singulars or is this just IDNs and variants?

SARMAD HUSSAIN So this is different from singular and plural similarity, which is a separate process. In scope, this is very focused on just visual string similarity.

CRYSTAL ONDO Okay, so that's going to be a separate discussion, separate sessions.

SARMAD HUSSAIN Exactly.

CRYSTAL ONDO Okay, thanks.

SARMAD HUSSAIN We have a couple of hands. Okay, so we have in the room Abdalmonem, and then … I'm sorry, I don't …

TAPANI TARVAINEN I'll introduce myself. Sorry, is it my turn? Okay, not yet.

SARMAD HUSSAIN Okay, we'll come back to you and then Nabil. So Abdalmonem, please.

ABDALMONEM GALILA Thank you, Dr. Sarmad. Actually, my question is related to the follow up to what you replied for Dr. [Anil, the Co-Chair]. You said that it is this stage, this evaluation, that will be at the prescreening for the one of the statuses of the applicant support program—one of the statuses. So it means that if one of the applicants pass this evaluation, and the other one tried to have one of the strings that could be similar to the other one that is already taken, it will not go for the prescreening process. The other one will be blocked.

SARMAD HUSSAIN Sorry. The prescreening tool is run after the applicant's applications are submitted, not before the application—during the process. And the prescreening is not for applicants. This is the tool which is used by the panel which is doing the string similarity evaluation. So once the applications are submitted, they go through the evaluation process. And one of the steps in the evaluation process is string similarity evaluation by an independent panel. And that panel uses this prescreening tool to determine what may be confusable strings. It reviews it manually and then creates the final set of confusing strings which are then put into contention sets.

ABDALMONEM GALILA

Yeah, it means they will collect all the requests, and then we'll evaluate that. But actually what make me confused is the word "prescreening," as there is another status for Applicant A [inaudible] B. It's called the prescreening. Thank you, Dr.

SARMAD HUSSAIN

Okay, thank you. So we can potentially use a different word but this prescreening is for the manual evaluation process by the panel. That is the context of how we are using it. I hope it's clear but we'll make sure we change the wording next time. Thank you.

TAPANI TARVEINEN

Okay, thank you. I'm Tapani Tarveinen from the Non-Commercial Stakeholder Group. A little question. I'm sorry. I haven't been following this very well in the past so this may have been answered before, but when you talk about visual similarity, I don't think that can be usefully defined without considering the language of the person looking at it. Different language speakers will confuse different strings. This is kind of obvious in case of different scripts because I say I don't understand Chinese so all Chinese scripts are confusable to me but it can also apply within, say, Latin script and its variants. So I'm not sure how you define this visual because visual actually happens in your brain but it looks similar to you. So, say, for a Dutchman, Y and IJ may look so similar that they can't tell them apart unless they think about it.

SARMAD HUSSAIN

Sure. So there is some discussion on this topic in the guidelines which we published, but I think there are some assumptions which are made, and they're document in the guidelines. For example, one is that the judgment is going to be based on somebody who's natively, for example, fluent in that script or reasonably fluent in that script. So, for example, if I don't know Chinese, the evaluation is not going to be done in the context of somebody who is not similar with a script. So that was one part of the question you had. The other part, of course, is that you can actually have people who are familiar with the script, but they actually use the script to write different languages. And there can be potentially some variation in how somebody can perceive two different characters because of the language they speak to write the script, rather than the same script. And those differences obviously are something which are going to be considered by the panel. It's not something which I guess will be out of scope. But on the other hand, I think, again, the assessment really needs to be based on the visual character of the string itself. And if the visual side of things are generally clear, even if they may be characters, for example, which are not used by a particular language, but visually, somebody who is familiar with that script can identify that difference, I think that is also something which the panel will take care of or take into consideration. So the answer to that question is that yes, the language does create some difference, but we would want to still do this analysis as close as possible to the visual distinctness or similarity of the strings. So that's a judgment which will eventually

have to be made by the panel. Thank you. So we have Nabil and then we have a couple of people in the queue. We want to go through the rest of the presentation as well, so if allowed, we will go to Nabil in the initial queue and then go through the next set of slides on data and then come back.

UNIDENTIFIED MALE [inaudible]

SARMAD HUSSAIN Ah, okay. Then we'll come back to you and then we can go on. So let's perhaps close the queue here and we'll come back to the questions again. Thank you.

NABIL BENAMAR Thank you. Sarmad. I just would like to add something here about the similarity or the confusion. So this goes beyond the concept of variant because, as we have defined for variants, we have the possibility to not distinguish between them, between two letters or labels. But for other confusion possibilities, it comes from, for example, joining some letters, and they might be really serious similar or identical to another separate letter. So this is one of the things that we have added to the definition of confusion. Thank you. And it's only related to the community using this script, so it's script-dependent. Thank you.

SARMAD HUSSAIN Thank you. Nabil. Ashley?

STEVE DELBIANCO Thank you. Steve DelBianco with the Business Constituency. Would the experts be allowed—

SARMAD HUSSAIN Sorry, there's a hand up by Ashley if we can go to her.

STEVE DELBIANCO Thank you. Will the expert panel be allowed to do evaluations of similarity even if it wasn't flagged by the software in the prescreening? Will they be allowed to consider things that weren't generated from the tools?

SARMAD HUSSAIN Yes. So as I said earlier, this is an assistive tool, and the string semantic evaluation panel is free to either ignore what is proposed or override and include what is not included.

STEVE DELBIANCO Great answer. So it's assistive, not restrictive. And Crystal made a question about singulars and plurals. So let's just stay focused on visual.

SARMAD HUSSAIN This is purely visual, and that's a separate discussion.

STEVE DELBIANCO I understand. And yet a long string that is an existing TLD coupled with an application which is the same string with an S on the end—is that not potentially visually confusing and similar? If you're ruling out that a simple addition of a letter would be visually confusing, then we have to go to another department to come up with singulars and plurals. But in general, does the addition of a single letter create confusion?

SARMAD HUSSAIN So basically there's no one answer to that, the reason being that I can add a letter in English and create a new word. Right? So you have A and then you add an S to a and it becomes "as," which is clearly distinct from the letter A versus the word "as." So just adding the letter S to a string may or may not actually create confusion. It will be eventually up to the panel to decide. And it also depends on which letter is being added. This really needs to be done in the context of the script. [It] may [not] be a layer of a language, though language obviously has some influence, as we just discussed.

STEVE DELBIANCO The bottom-line question is, would the panel be allowed to say that the addition of an S to a long string is visually confusing? Would they be allowed to conclude that or does it get punted to a different process?

SARMAD HUSSAIN So yes, if the panel considers something to be visually confusing, it is in the remit, and that still does not mean that it will not be again reconsidered in the singular/plural part. Okay, so then I think we move on. We have a question, yeah?

ASHLEY ROBERTS Thank you. Just with regard to the evaluation tool, just to clarify, I believe you're looking at trying to make that available for applicants to use as well. Is that right?

SARMAD HUSSAIN Not yet. That's been a request and we are looking at it. But let's look at the tool. We'll come back to your question. One of the things we are considering is, once … Since we haven't developed the tool, we can't answer that question. But when we do develop the tool, we are looking at the possibility that we can release it open source so that people can just install it and use it on their end. But there's probably not a likely chance that we will make it available online for the applicants. This is currently being designed for the panels, but we can release the tool for people to use and install and use. Let's move on in the interest of time, and I'll request Michael Bauland to take us through the data collection part of this first string similarity evaluation.

MICHAEL BAULAND

Yes, thanks. My name is Michael, and as Sarmad said, I currently managing the data collection part as the data is defined by the root zone LGR because all code points that are in the repertoire are basically allowed for TLDs, and these repertoires are within 25 scripts, soon to be 26 scripts, with Thaana being just added to the root zone LGR. And all these repertoires need to get analyzed and this is too much data for a single person or even a small team to look at. And also this would require the team to have knowledge of all the scripts. So for that reason, the task is being delegated to experts within each of the scripts, and each of these experts has a scope of work. They have to compare their repertoire, one, with ASCII characters (so all letters A to Z, both small and capital, have to be compared to the repertoire) and then there's of course the in-script comparison. So all code points within the script have to be compared to each other. Then there's potentially a cross-script comparison that depends on the actual script because some scripts belong to the same family and therefore have some similarities and thus require the experts to also compare their repertoire with the repertoires of the scripts within the same family. And finally they have to look at miscellaneous comparisons—for example, capitalized elements within the script, if the script supports capitalization, conjuncts and compounded characters, and simple shapes like O and L, lines, circles, curves—and also to look at underlined versions. This was included because if you look at a domain name on a web page as a link, usually browsers underline those links, and the underlining can hide parts of the character, which may make a certain character look exactly

the same as another one, even if, without underlining, they would be quite distinct. Next slide, please. So the experts were asked to categorize the comparison into five categories. One is the identical or near-identical pairs, and this also includes all already-defined variants within the root zone LGR. Two are the highly confusable pairs. These are strongly similar or near-homographs. Then we have category three which are the similar pairs. And four and five are not looked at explicitly, but five is just a majority of pairs, everything which is distinct, and 4 is weakly similar. And this is only to be included to state clearly that this pair has been looked at and it was decided that it's not within categories 1 to 3. Next slide please. So I'll now show some more details for each of the categories and provide also some of the examples that have been created by the scripts experts. Because I won't show all 25 scripts, I tried to use a selection of scripts that quite diverse and will show you some data that has been collected so far. So category one. These are the identical or near identical plus variants. So all existing variants automatically belong to this category. The experts do not have to add them manually. And then this also includes other cases of identical characters, for example due to capitalization, which was out of scope for the variant determination, and also near-identicals which are easy to confuse even if you see those characters side-by-side. Next slide please. So here you can see some examples from Hebrew, Latin and Bengali. I won't read them out because other than “testi,” I don't know how to pronounce that. And as you can see in the Latin sample, those strings are easy to distinguish if you look at them this way. But if

you imagine an underlining due to URL highlighting, this might completely hide the underline below the "I" and thus make the first and second string actually look exactly the same. Next slide please. Then we have category two, the highly confusables. These need to be confused by a large majority of attentive users when not looking at the cases side-by-side and you see two examples from Ethiopic and from Latin. I'll just leave it here for … Okay, I just wanted to give the audience a few things to look at the string. And next slide, please. And then category three. These are the similar cases these need to be confused by a simple majority of attentive users. So the distinction to the previous case is that this is just a simple majority, whereas category two needs to be a vast majority. And this also includes a special case for mainly the Latin script, namely the fallback replacements due to Latin script, also including ASCII. There are several mostly historic reasons where people using non - ASCII Latin characters would replace the characters, for example with an accent by the ASCII version if such a character was not available on a typewriter or keyboard. And again, here are two examples from Latin and Bengali that would fall into this category. three. I'll wait again a few seconds for you to look at those examples. Next slide, please. Then we have category four. These are only distantly similar. So they are kind of similar but not really confusable or just to a minority of attentive users. And as I said before, experts are not actively looking for these cases, but they should note them just to make sure that people looking at the data can see that this case has been looked at and not been forgotten. But the decision was that it's not similar enough to be in one of the

first three categories. And here we have two examples from Hebrew and Ethiopic. Okay, next slide please. And then finally the distinct characters, which is a majority of all pairs. For example, you could see an I and M, for example. These are certainly not confusable at all. They are completely distinct. And there is no examples here because almost all pairs fall in this category. Next slide, please. So what is the status of this data collection project? We are almost done. The majority of the scripts is currently preparing for public comment. We have a few scripts where the internal review is in progress. And two scripts are a bit behind. There's one which is in preparation. This is Thaana, as this has only been recently added to the root zone. And then we have one where the expert is still working on the initial data. And this is the case for Myanmar, as there has been some issues that caused some delay in the work. Next slide, please. Yeah, that's it. Any questions?

SARMAD HUSSAIN Thank you, Michael. So we'll take a couple of questions, but because we have to get through the rest of the presentation, we'll take the remaining at the end of the presentation. So two questions and then we'll move on please.

[TAPANI TARVEINEN] Just a quick clarification. You mentioned the fallback replacement in Latin script. Does that Include cases where the conventional fallback is multiple characters, say German umlaut versus ae?

MICHAEL BAULAND No, this is not included here because this is purely on visual similarity. And while this is a common case to do such a replacement, it's visually not similar at all.

SARMAD HUSSAIN Abdalmonem?

ABDALMONEM GALILA Thank you, Dr. Sarmad. Actually you said these are five categories, which is good. But look, if one of the words is part of the other word (for example, God and Godi), this is one of the issues. For example Google: G-O-O-G-L-E. Maybe it is G-O-O-O-L-E or whatever. How could you handle this? Or it shouldn't be handled or it will be considered as a different word?

MICHAEL BAULAND So this data collection part is not really on a label level or word level. We are just comparing code points or conjuncts to other code points. So this data just looks as these pairs, and then the software tool will use this data to apply this to actual labels and strings.

SARMAD HUSSAIN Yeah. So the determination of confusability is a separate step. This is just the data which will feed into that discussion.

ABDALMONEM GALILA Why is it this? Why do [you and the regulator] try to validate mobile numbers? We have one of the functions inside the database that

checks similarity. When it is more than 50%, we'll will have an issue here. We need to check. We found something: one number exchange[d for] the other number. That's what I am talking about. Thank you very much.

SARMAD HUSSAIN

Thank you. So then let's move on to the next part of this presentation. I'll hand it to Julien to take us through the tool which is being developed.

JULIEN BERNARD

Thank you so much. I'm Julian Bernard. So I'll talk about the SSE evaluation tool. So the flow of this tool is . So we have the SSE tool that is configured with the SSE data, the one provided from which Michael talked about. And it's also configured with the various names; TLDS, ccTLDs, block names and all of the kind of names that we want to compare with the applied-for strings. So this tool will receive the list of all the applied-for string which is the TLD application management system and generate some system reports based on the data. Then some panelists will be able to look at these reports. And if a panelist thinks he does not have sufficient data for a decision, he will be able to run again with a subset of the applied-for strings and threshold value. We will talk about that later. So the tool with this subset of applied-for strings will be able to generate the custom reports. Then the panelists will be able to look at this custom report as well as the system reports. And then the panelists will be able to edit the outcome of the tool to add, remove things, add comments and then provide the output to the

panel manager. The panel manager will then gather all the reports and assign them inside the final report. This final report will then be provided back to TAMS, which will handle the determination. So for instance, it decides the applied-for strings are similar to applied-for strings, and therefore they are in contention and they reject the strings that are similar to block names, and so on. Just checking if I did not forget anything right here. So just to confirm, since we had questions about that, the tool is only a guide, and the panelists will make their own decision with the help of the tool output, either the [system one] or any custom report they will get based on what they provided as input. Next slide, please. So the tool can be accessed at ICANN.org with an ICANN account for the panelists, and it will provide the automated string similarity analysis with the goal to optimize the manual review. The string similarity data described by Michael in the previous slide will be incorporated in the tool, and the applied-for string provided as input will be compared to the list of other strings as described by Sarmad—so all of the applied-for strings, existing gTLD or ccTLDs, requested IDN ccTLDs, the previous round, in-process applied-for gTLDs, two-character ASCII, and block names. So basically the tool will perform this in three steps. First it will detect the potentially confusable labels. Then it will compute the visual distance between those labels, and when labels get scores that exceed the threshold, they end up being included in the tool final report for manual review by the panel. Next slide, please. So about the threshold, the threshold is a configurable value that will determine the comparison sensitivity. So if we have a similarity score between

a pair of strings that is below the threshold, then the tool will report the variant set with the most similarity for this pair of strings. And regarding the outputs, we have two kinds of report; the system report that is generally based on all the applied-for strings (this report will be generated using some predetermined threshold values), and then we have the custom reports that are generated based on the subset of applied-for strings and triggered by the panelists. And along with the subset of applied-for strings, the panelists can provide a custom threshold value. And those reports will be available in CSV and HTML format; CSV to help panelists edit the data, and HTML will provide the report in a way that it will be displayed without any issue that you might encounter with any tool panelists will use to open a CSV file. Next slide, please. So the basic mechanism for the tool is that, based on the similarity data files, it will calculate if any applied-for string or its variants is the same or similar to other in- process, applied-for strings, existing TLDs, requested IDN ccTLDs two-characters ASCII, blocked names or their variants. So first the similar strings are grouped in potential contention sets, and for each potential contention set, a score is computed between each string variant pair. Then the report for the panelist to do the final review will contain the pair of potentially similar strings with a minimum score below a certain threshold. So when we compute the score between string variant pairs, we will have multiple scores depending on the string and or variant. We compare them together and for one pair of strings, we take the one with the lowest score. And the metric to calculate the similarity score will be determined after the data analysis. Next slide, please.

So the output report will basically contain this list of elements, some contention set, this unique ID, the label, and the application ID of the label that are in contention. If label 2 is not an applied-for string, then the source of label 2 and then the variants for the labels end up having the most similarity in the comparison, then you will have a score. Some [inaudible] we only have code points with a similarity level below or equal to 3, and other fields are elements that the panelists can use to put comments, notes or anything helpful for the final report. Next slide, please. So this is an example of how the report will look like. For this theoretical example, we have block name which is example and some similar strings that will end up in the report—for instance, “m” replaced by “rn,” or “l” replaced by an “I”, which is already similar to an “l” when uppercase and also “I” with grave accent. And that finishes my part. Any questions?

SARMAD HUSSAIN I don't see any hands in the room and in the chat. So then let's move on to David for just a quick discussion on timeline and then we'll come back for any more questions. David, please.

DAVID WAGHALTER Thank you. I'm David Waghalter. I'm a Product Manager with the Engineering and IT Group at ICANN, and we're working on this tool that's just been described. This is our development timeline that you're seeing here. Currently we actually are in development. We do have most of our requirements complete for this round or for this stage. We are finalizing them, kind of tweaking and fine -tuning

as necessary, and we'll continue development through the month of April. At that point we're going to have an internal UAT on what we developed for our MVP, Most Valuable Product, and from that we intend to unearth some additional requirements, any fixes that might be necessary once that's been evaluated internally. Then in June we'll develop requirements out of that outcome and we'll continue development through the summer with a goal to have our QA, internal hardening work, and then our release in September. Okay, any questions on that?

SARMAD HUSSAIN Thank you, David. Okay, so then that brings us to the end of the presentation. And let's open it up for any questions you may have. Thank you. Or comments.

VINNY ADJIBI Good morning. Thank you for the presentation. I wanted to ask, is the data used for the evaluation going to be made available?

SARMAD HUSSAIN Yes. And if I may request you, please introduce yourself.

VINNY ADJIBI Vinny Adjibi. I'm a NextGen participant.

SARMAD HUSSAIN Thank you. So yes, as we shared, we are planning to publish the data which we've received from experts for public comment

feedback. And based on that, then of course the input we receive from the community, we will finalize the data which will eventually be used by the tool. We should have this public comment come out in, I think, the next two to three weeks. I'll just give another minute to everybody to see if they have any more questions. While you're thinking, just to again reiterate, that this tool is still being developed and as this is still in process, at this time, we cannot make sure that the tool is actually going to be available for the community. This is really being developed for the panel. Once we have the tool, then at that point we can make any determinations on broader availability. Yes, question, please?

STEVE DELBIANCO

Thank you. Steve DelBianco with the Business Constituency. Yesterday the BC held its meeting here and got presentations from vendors that assist registrars at blocking new registrations of domain names that are intended to defraud and deceive, often with similarity and homoglyph attacks. They have indicated a dramatic increase in their effectiveness by the use of some GenAI LLMs to do the analysis, even with cultural context. But apparently they rely heavily on prior history of the registrant, which would indicate perhaps an intent to deceive. None of that would be valid here, but I wonder how broadly have you explored incorporating AI into the tool. And of course the experts might choose to use a chatbot on their own to compare things. But if you said it earlier, I apologize for missing it, but are you using AI in any way in the tool? Thank you.

SARMAD HUSSAIN So the AI is not being used in the tool. It's purely based on calculations based on the ranking of similarity provided by the experts. But in at least one script, I know some of the data which was produced had some machine learning, I guess, employed to identify what similar code points are. Specifically that was done for Han script in Chinese just because there are so many different characters. So the expert who was actually determining the confusability of different characters actually employed some automated machine learning techniques to do some shortlisting and then, based on that, made their suggestions. And when we publish our data for public comment, we'll share those details as part of the process.

STEVE DELBIANCO I'm sorry. By AI, I was being specific. I don't mean machine learning. I mean GenAI based on a large language model. Given the world we live in and our technical focus, you have to accept that by next September, if we release a tool that has no acknowledgment of the value that AI could provide at doing the analysis, we're going to look foolish. And the rest of us will use AI anyway to evaluate decisions that came out of a tool. This is 2025, your ICANN. You need to at least look into using a free LLM as part of the process.

SARMAD HUSSAIN Sure, we'll take that feedback. Thank you. Yes?

ABDALMONEM GALILA: Just a comment. Actually AI depends on some data that already is there. Maybe it is big data. This is database. It's data itself. But I think for LGR and similarity strings, it depends on code points. So I think it is somehow different. If AI could depend on code points (and I doubt it), I think it will be good. Otherwise I think it will not be the good option for that.

SARMAD HUSSAIN: Thank you, Abdalmonem. In any case, we will take that comment back and we'll take a look at it. We have another hand up. Vanda, please go ahead in the Zoom room.

VANDA SCARTEZINI: Hello. I just want to inform you that in our registry if you want to register an IDN domain with similar … For example, the letter O in ASCII and first letter O in Armenian is similar. For your eyes, it's similar but you can't register an IDN domain name using another font—for example, [Orege]. If the first O will be on ASCII, you can't register this domain name. I just want to inform you. Maybe it is [of] help to register phishing sites with confusing similarity.

SARMAD HUSSAIN: Thank you, Vanda. Okay, so we are on time. If there are no more questions or comments (I don't see any hands up) then thank you all very much for attending. We'd like to thank the panelists as well for their contributions, and we'll close the session. Thank you.

[END OF TRANSCRIPTION]