DEVAN REED: Good morning, good afternoon, and good evening. This is Devan Reed for the recording. Welcome to the Latin Script Diacritics PDP Call taking place on Wednesday, 23 April, 2025 at 13:15 UTC. We do have apologies from Sebastien Ducos and Amadeu Abril. Statements of interest must be kept up to date. If anyone has any updates to share, please raise your hand or speak up now. If you need assistance updating your statements of interest, please email the GNSO Secretariat.

All documentation and information can be found in the Latin Script Diacritics wiki space. Recordings will be posted shortly after the end of the call. Please remember to state your name before speaking for the transcript. And please note, all chat sessions are being archived. As a reminder, participation in ICANN, including this session, is governed by the ICANN Expected Standards of Behavior and the ICANN Community Anti-Harassment Policy. Thank you, and over to you, Chair. Michael, please begin.

MICHAEL BAULAND: Thanks. First of all, I just want to mention that I got a small cold or flu last weekend, so my voice might still sound a bit strange, but yeah. The good thing is that we have remote session here, so I don't have to wear a mask to not infect you. That's quite an advantage of the remote thing.

So, this is the agenda we want to go through to that day. We have quite a few topics here. We start with the welcome. We start with the recap of the previous meeting. Then we will have our guest speaker, Sarmad,

who will provide an introduction of the String Similarity Review Process. And then we will continue with Charter Question 3.

Yeah, so we'll start with a recap. The key outcomes of the previous meeting or the previous meetings is that we agreed to consider all TLDs, whether they already exist or are only going to be applied for. This will make no difference for our policies. We looked at the ccTLD Fast Track process and decided there was little relevance for us that's not already covered elsewhere. So, we decided to leave it at this and potentially come back in the end should there be anything in the Fast Track that's not covered, for example, by the EPDP IDN recommendations.

We also saw the list of Unicode characters that are likely in scope, which was generated kindly by Mark, and he will say a few more words about that in the later part of the meeting. And for the key action items, we had the open question whether the base ASCII TLD would be a requirement or whether only diacritics would be a possibility too. And we discussed this within leadership and we will present the solution to you later in this meeting. As mentioned, the analysis report for the ASCII Unicode characters has been uploaded to the wiki page.

And we also talked about the idea of developing examples of edge cases for stress tests. And while we think that's a very good idea, I think it's a bit too early to do this now as we have hardly any recommendations yet. But we will definitely come back to this when we are nearing the finalization of this PDP to have then the recommendation tested against some examples and possible edge cases.

Yeah, some good news. Our GNSO Council Liaison, Prudence, presented the project plan to the GNSO Council and she did it in such a good way that there was no objection and it was approved. Thanks to Prudence for that. And here I will hand over to Mark for the work he's been doing. Thanks. Over to you, Mark.

MARK WILLIAM DATYSGELD: And hello, everyone. Thank you for introducing the subject, Michael. So, you might have seen circulated this particular piece of work. And let me contextualize it a bit. So, we had a few options to go with this, but since we were heading towards a specific interpretation of what the scope of the work is, I settled on doing this in a more programmatic way. We could have outsourced this or asked for a report or something, but I'm proposing this method in which we just are leveraging the power of Python and Unicode to solve our problems for us.

I coded a tool. I released it in open source. It's on the link right now. And the idea is even if we change the scope slightly for some reason, it's fine. Anybody can just tweak that and add to the existing functions that generate this report. So, basically, we're cutting a lot of time away from the process, right? We have already the calculations in place. We have already the generation of the report. And what that does is it adds a lot of transparency to what we're doing. There's not a lot of, hey, trust me, this is what's within scope. It's, hey, this is in scope. And I can prove it to you.

So, hopefully in terms of us showing to the community what we're doing and why we're doing it and what we're considering, this will make

things a lot more straightforward. Because if somebody asks a member of the group why, the answer is because mathematics and analysis tells us why. So, essentially, what we want to do with this now is we have our list of, let's say, within scope characters. It doesn't mean that they are necessarily final, but within our current understanding, we do have a list of what is final for now. And we can work on top of that.

And I don't know if everybody had the chance of having just a look, like a very quick one. And just in case it hasn't been the case, I will go over super, super fast. Just to give you context of what we're working with. And as you Saewon is helping me. Thank you so much. So, we have the characters with one diacritic mark table. This is sort of the over table and the ones we expect to work with what is within our scope. So, we have the character, the decomposed combination, and then the actual names within Unicode of what each character or operator is doing. So, we can point out very clearly to applicants and to the community and anybody who asks, like, hey, is this within scope? You can just point to the table and say, hey, yes, because it's this plus this. And I can prove it to you because this base plus this diacritic.

And if you go down, Saewon, I think it's towards the very end, the multiple compositions, we do have a few cases—right there, thank you so much—of characters with two diacritic marks that fall within scope. Those are 30, as you can see, numbered. They're more unusual, but they are perfectly within the realm of what we set out for this work. We might want to have maybe a more specialized look into these, just to make sure, and talk to Sarmad more deeply a little bit about these ones, but that's all good. And yeah, Bill has a very good point. At least

one of them for Vietnamese is actually pretty common. Yeah, you make a great point, Bill.

And finally, at the end of the document, if you scroll down just a tiny bit more, our fearless leader, Michael, pointed out that some cases are within the Latin RZ-LRG, but are not derived as mathematically, let's say, as the others. So, these are the sort of the inclusions that are not exceptions. They're things that made it into the RZ-LRG and our additional cases. They are the only exceptions that they are not mathematically derived. They are sort of put into place. And that's my entire explanation. Satish, please.

SATISH BABU: Thanks very much, Mark. I think, personally speaking, I greatly appreciate the work that has gone into this. And I think it's an excellent contribution from the community. It is open source, and I can't think of a better model of doing something like this. But I do have a question on what makes it official, in the sense of the Root Zone LGR tool is ICANN's own tool. So, that makes it automatically official. Now, this is a community developed tool. It may be completely perfect in every way, but is there any way that we can, so if a third party ask, okay, what is the source of legitimacy of this? Has somebody certified it? What answer do we give? Thanks.

MARK WILLIAM DATYSGELD: Thank you, Satish. The answer is self-contained, I think, because yes, it is meant to be entirely community driven, because exactly it's coming from the community. It's what we as a group are approving. And my

idea is that once this gets reviewed by the community, and this gets published into our reports, and this gets included in our normal stream of documents for this group, then it's when it gets approved. So, it exists as a reference for all of us for now. We can have a look at it, and we can see if there's any need for revisions.

And the moment we send out our public report, we just pack the final version of this report, as we understand it, the link for the two, and there you go. It becomes an outcome of the group's work. And what this also facilitates is that in next PDPs, we will be able to encourage people to use a similar solution and maybe speed up the work on the different things that we have been discussing. So, this is also forward looking, right? We are trying to see if we generate a repeatable model to accomplish all the things that we have been discussing and actually solve this entire problem, if not within this PDP, at least in the very near future. Thank you.

MICHAEL BAULAND: Maybe as an addition to that explanation, I think this is very great. Thanks, Mark, for all the work you put into this. And it's very helpful, but we should not consider this to be authoritative. So, the authoritative decision, what is in scope or not, is the actual recommendation as we have written it down, namely that the Unicode table will list such a character as a combination of a base character with diacritic character. So, we can provide this list, but we should make sure that people are not taking this as the authoritative source. It's just a tool, a helping thing we provide for convenience.

Okay, I saw a question by Bridget. I think Saewon already answered that one. Any further questions regarding Mark's work? If not, then we can come back to the slides now. And thanks again, Mark, for this great work, and also for making the code public so everybody can look at it and potentially adjust it if that were necessary. But I think it's quite good as it is right now.

So, now we come back to the question we talked about last time. Namely, we already decided that we would not want to restrict this to only two TLDs, but it should also be possible to use our recommendations for three or more TLDs should that become necessary. The only corresponding question was whether the base ASCII TLD would have to be in the set of these TLDs, or whether it's also possible to just have our policy apply to diacritic TLDs, like the second and third line of the table, test and test with DXN.

And we talked about this in leadership, and we basically came up with three possibilities. And the first one is just to stay within the current scope, and that's it. The second one is to stay in the scope, but similar to those characters we found which somehow look similar to diacritics, but are not actually diacritics, we decided to not extend the scope, but to include recommendations for potential future work. And this was the second solution to do the same thing here, not to extend the scope, but to say this could be future work to have recommendations to also allow not using the basic ASCII version. And the third solution would be to submit a project change request to the GNSO Council and try to get our scope extended to include this.

And we thought, we discussed this quite a while, and while many of us think, or even all of us think it makes sense to have rules for cases without ASCII, we don't want to start changing the scope right now. This would really endanger the timeline of the PDP, and we want to have a very narrow-scoped, quick PDP where we deal with the scope we have been given in a good way, and then open the opportunity to look at all the trauma cases or similar cases we discovered during our PDP in some future work, maybe another PDP, maybe some other process, but that's out of scope here too. But we will list these and remember them and write them down in the end.

And we think this is the best way to go forward to have a clean and good solution that can then later be extended to other cases as well. Because if we now start to extend the scope here and there, it can easily lead to a situation where the PDP will run for three or four years because there are always new things that are discovered that would also be worthwhile to discuss and make rules for, but are not currently within scope. So, unless anybody here has any objections to this approach, we suggest to do it this way. So, if you have objections, please raise your hand. Okay, I see confirmation from Satish and no objection for--

JUSTINE CHEW: Sorry, Michael. This is Justine. I can't find my raise button.

MICHAEL BAULAND: That's all right. Justine, please.

JUSTINE CHEW: Yeah, sorry to interject. I am having a little bit of trouble understanding the implications of Option 2.

MICHAEL BAULAND: Okay. The implication, well, for one, the recommendations we do are exactly the same as for Option 1, I mean the actual policy. We will just mention in the end of the PDP, I'm not totally sure how this would work out on a technical level. Maybe Steve can chime in here, but we would just list cases we've discovered that are worthwhile to cover, but that are not within the scope. And we don't want to extend the scope, we want to list those cases. And most likely our rules can just be used for both cases too, but we won't decide that here in the PDP. So, we just do suggestions for future work. Or, Sarmad, maybe you can help with that. Thanks.

SARMAD HUSSAIN: Yes. I think if the PDP, this PDP is recommending something, they would probably need to analyse those cases before they can recommend, unless you word it in a way that you actually do not recommend, but perhaps suggest exploring other areas, which potentially could be done without analysing. But if you start analysing, of course, that goes back into your original argument that means that more time will be needed just raising that difference for your consideration. Thank you. This is Sarmad.

MICHAEL BAULAND: Thanks, Sarmad. Good point. We don't want to start an in-detail review there, so maybe the second option you mentioned is better. Steve, please.

STEVE CHAN: Thanks, Michael. This is Steve from staff. And helpful clarification from Sarmad. Yeah, if this group wants to have anything result in implementation, then it needs to make sure it has all the analysis, as you mentioned. What I wanted to point out is the procedural way in which Option 2 could occur. Put the link in chat. In the PDP manual, I think it's the PDP manual, yep. On page six and seven, it talks about acceptable outputs or allowable outputs from a PDP. And the very last item in the list of potential outputs, which includes things like consensus policies, of course, and best practices and a bunch of other things, it actually mentions at the very end recommendations on future policy development activities.

So, I think that would be something where this could fit, but like where the PDP itself is not going to do the robust analysis that Sarmad rightly points out would be needed if there's a recommendation that requires implementation. But it could recommend to the Council that there be additional work considered. So, I think that would be the mechanism if this group actually ends up going down Option 2 for this and anything else that comes up that is not currently within the scope. Thanks.

MICHAEL BAULAND: Thanks, Steve, for that. I think that's the best way forward and we can look at the details how exactly we trace that when we are nearing the

end and have to write the official document. Thanks. So yeah, I see no general objection with this approach, so we can continue to the next slide, please.

This is basically something we have already seen last week, just a diagram to easily see what is in scope, what we will be doing and listing the key outcomes. The only thing we added is the first line in the key outcomes in this dotted or dashed box, namely that we include base ASCII as a requirement for the delegation and operation of gTLDs in the context of this PDP.

Then we have the recap regarding the Charter Question 2, if a solution is needed to the issue, are any of the elements of the ccTLD Fast Track process transferable? And as already mentioned earlier in the overview, we decided that at the moment we don't think this is necessary because most or even all of the things worthwhile for us in the Fast Track process are also covered, for example, in the IDN EPDP paper.

So, this is a summary chart we are going to develop through the course of this PDP. And we've filled in the first line for the ccTLD Fast Track process. And as mentioned, we don't really see transferable elements here. It was mentioned that the Fast Track process has some same entity principle requirements, which of course we also want to have. But since these are discussed and described in much more detail in the IDN EPDP, we will more likely take those recommendations and reuse them rather than trying to get this out of the Fast Track process. So that's why we still put an A there.

Are there any questions regarding what we've been discussing in the last meeting and which has been summarized now? If not, I would like to hand over to Sarmad for his introduction into the String Similarity Review. Sarmad, please.

SARMAD HUSSAIN: Thank you, Michael. This is Sarmad. So, this is a presentation which we had done earlier as well at ICANN82, but we are trying to reuse the material to the extent possible for consistency as well as, but happy to answer any questions you have. If you go to the next slide, please.

So basically, what this presentation does is provides an overview of the policy as well as based on the policy development which was done, the need for next steps and what are some of the next steps which we're taking care of. It includes actual data collection and tool development, and then we'll share some timelines as well.

So basically, string similarity requirement comes from the 2012 Round of New gTLDs and also from the IDN ccTLD Fast Track process, which is, I guess, in some ways earlier than the 2012 round because that was published in 2009, though obviously not directly relevant here. For the next new gTLD round SubPro recommendations affirmed that the string similarity, which was something which was, I guess, taken up in 2012 round should be a continued check on the new strings being applied for under the New gTLD Next Round. And it affirmed that we will use the same visual standard for the Next Round as well as was used in the previous 2012 round.

And the visual standard which was defined in the policy and has been repeated here as well is that similar means strings so similar that they create a probability of user confusion if more than one of the strings is delegated into the root zone. So basically, what this means is that we're not looking for just a possibility of two strings being confused, but probability of two strings being confused, meaning that there is a reasonable likelihood of that happening and not a low likelihood.

And the other thing which is important to note is that it is a visual similarity standard. So, it does not extend to basically phonetic similarity or semantic similarity, meaning or sound, but purely visual in nature. So those basically are the definitions which come through the SubPro policy. Of course, saying that there should be a probability of user confusion rather than, of course, a possibility then obviously leaves some level some detail to the interpretation to the or for the panel.

In IDN EPDP Phase 1, which builds on top of SubPro for IDNs, there are additional three recommendations. Recommendation 4.1, which identifies the scope of string comparisons. There's more detail on that in the next slide. We'll talk about that in a bit. Then Recommendation 4.2 states that the decision by string on this really needs to be through a manual process, not using like an automated tool. And that needs to be done through a panel of experts, which we call string similarity. We actually change the terminology now or evolving it and calling it String Similarity Evaluation Panel and not a review panel.

And then Recommendation 4.3 states that even though there are many comparisons which could be done, the string similarity review panel or evaluation panel can—based on string similarity review guidelines—skip

some comparisons or do some additional comparisons. So, they have been empowered to really make the final call on what is similar and what is not. And more about that in the next slide. Let's move on. But maybe let me stop here and see whether you have any questions so far.

MICHAEL BAULAND: There's one in the chat from Tapani. Strings that are so similar that they create a probability is rather imprecise. What are your impressions by that comment?

SARMAD HUSSAIN: Yes. So, thank you, Tapani. But that is really the policy language, which we are working with and we'll share with you some details on how we are parsing through that and creating something which is, I guess, more precise. And I will come back to your comment again once we've talked a little more about this. If that's okay, then we can proceed to this slide.

So, this is the scope of comparisons. And basically, what you see is that when you have applied for string, which is at the top of this table, top right, the applied for string can have a primary string, which is the main string being applied for. And then it can actually have allocatable variants and blocked variants. Allocatable variants are those which an applicant can apply for in addition to the primary. Blocked labels are those which are variants of the primary as defined by the Root Zone LGR, but they cannot be applied for by the applicant.

Similarly, on the left-hand side, on the first column here, this string, which is being applied for will eventually need to be compared with the

whole set of other strings. It needs to be compared with existing gTLDs, and the primary of the existing gTLD, all of its allocatable variants and all of its blocked variants. Similarly, the gTLD string applied for in the previous round, but still in process, if there are still any left, they would all also be part of this comparisons. In addition, we have the existing ccTLDs, requested IDN ccTLDs, which are in process, other applied for gTLDs, blocked names, any two character ASCII.

So, basically the categories with which the applied for string is going to be compared with, we will compare the primary, allocatable and blocked variants of the applied for string with these categories, their primary strings, all allocatable strings and all blocked strings. So that creates nine ways of comparison. And as you can see, only one of them, which is blocked versus blocked is out of scope. And otherwise, all the others are in scope.

And then there are these two grayed out rows. Those are areas where I guess the panel has more flexibility where they could actually even skip some of the analysis. For example, one example is that if they are comparing a Chinese string with an Arabic string, there is a likely chance that the panel may determine that those two scripts are so different that there is a low chance of similarity. And therefore, they may not want to compare all the strings with all the strings and say that we'll skip some of those because just that pair of scripts are too distinct with each other. But that's a decision which the panel has to make.

Any questions on this, on the scope of string comparisons? Then move on, please. Justine, please.

JUSTINE CHEW: Yeah. Thanks, Sarmad. Sorry, this is Justine. Just to add, the panel can determine what to omit, but I believe they also have to provide a rationale for what they are omitting.

SARMAD HUSSAIN: Yes, true. If they omit, they will document the rationale. Actually, they would need to document the rationale on all of their decisions. So, yes. Moving on then. So, to guide this process, the policy suggests that we also develop some string similarity guidelines. These guidelines are in some ways to guide the panel for to do their work just to make sure that there are obviously some guardrails to this work, which are defined by these guidelines.

There was an initial, very early version of these guidelines put in for public comment. And we received more feedback. We're based on that developing some more data and tools. That data and tool and that mechanism was actually suggested in those early version of the guidelines. Now that once we have the data and some, I guess, a clearer definition of the tool, we will obviously go back and update the guidelines and make them, I guess, more precise and bring them back for public comment again. But currently, we have been doing more of the homework, which is needed. And we're very close to now closing that homework. And based on that, we aim to bring these guidelines, update these guidelines and bring them back for community feedback, perhaps in the next half of the year, calendar year.

So, that's sort of some bit of a background. One of the things which we actually identified in the guidelines is that the number of comparisons is now very large. As is without variants, the number of comparisons was quite large in the 2012 round as well. And that had caused the panel to take up quite a bit of time to do the analysis. This was they were around 1900 strings or something all needed to be compared against 1900 other strings, and so on. So, that was the scope of last comparison.

Now, we'll have these new strings which will be applied for and they'll need to not only these strings, but their variants will need to be compared with existing strings, not only existing strings, but also their variants and so on. So, the scope becomes very large. And so, there is probably need to, we felt there was need to do provide some assistive technology tools to the panel to help them go through this more, I guess, in a timely fashion.

So, what we have been working on and what we had proposed in the guidelines as well was that we develop this pre-screening tool. Again, this does not bypass the manual oversight of the panel, but it just helps guide the process better. And it is designed in a way that it has what we're doing is we're collecting similarity data from script experts for all the scripts in the Root Zone LGR. And we are actually developing a tool which uses this categorization of similarity by the experts to give hints to the panel on which two labels could be similar or which two labels could be less similar. And then eventually the panel really has to decide how to use the output of this tool or not even use that output of the tool and move forward and determine the final analysis in more detail moving on.

Please feel free to raise a hand in case you want to ask a question. I'm more than happy to stop and respond to questions as we discuss the particular topic. So, similarity data collection is ongoing work. This is based on the repertoire, which includes 25 scripts based on Root Rone LGR5 because those are the only characters which are allowed through policy to create basically strings for gTLD applications. So, anything out of it will obviously not be applied for. So, we don't need to worry about similarity cases for those characters.

We are also currently working on Thana script, which is a 26 script. So that will be added into the scope as well as Root Zone LGR Version 6 is compiled. So that will be 26 scripts in total. All the characters in all these 26 scripts will be analyzed by the experts and they will identify similarity based on whether any of those characters is similar to any ASCII character or any other characters within that script or any other characters across different scripts, especially those scripts which are from the same family.

When we say same family, basically Latin, Greek, Cyrillic, for example, are from a similar family. The new Brahmi scripts are similar and from the similar family and so on. So, we have these families of scripts which are closer to each other. And so, it is useful to look at those cross-script variants more carefully than unrelated scripts. And then we obviously want to look where it's applicable.

We want to look at uppercase similarities. In many scripts when two characters join, they create a new shape. So those are called conjunct consonants. So, we want to look at conjunct consonants and compound characters. We also want to look at similar shapes across all the scripts.

So, things like just open circle like an O or a vertical line. These kind of simple shapes are sometimes present accidentally across many different scripts, even if they're not related. And then we are also considering underlined versions of characters because URLs are many times underlined. So those are some of the-- That's the scope of the data collection.

And this is how we are subdividing similarity. So, this is going back to the question which was asked by Tapani on how do we define probability over possibility. So, what we've done, again, this is assistive mechanism, right? This is not really-- Eventually the panel has the final decision. But just to assist them, we are going to be defining the different similarity of characters we are determining from experts into five categories. These categories include characters which are totally identical or near identical, so not easily distinguishable. Some of these are already captured through the variant process, but even in case there were some which were not captured through the variant process, they'll be captured in the string similarity process at level one.

Things which are highly confusable and so strongly similar, but not identical or indistinguishable occur, we requested experts to categorize at level 2. Characters which are similar are categorized at level 3, then distantly similar or weakly similar, meaning they could be similar but not really convincing are level 4, and then characters which are totally distinct are level 5. And what we are saying is that if we define this in these five levels, anything which is level 3 and above we would consider similar. And then in level 4 and level 5 may actually be not similar but 3 and 4 obviously will provide some gray areas.

And not only that, it is also, I guess, important to note that when we do this categorization, this categorization is done at character level, eventually panel has to review strings not characters. So, the tool will obviously analyze this at a character-by-character level but eventually panel has to make a determination not at a character-by-character level but at a string level, and that determination can obviously those two determinations can deviate.

So again, this is, as I said, an assistive tool just to allow the panel to, for example, maybe shortlist things which are-- So, I guess, the tool can help the panel shortlist strings in maybe 2 or 3 categories. One category would be things which are totally similar to each other and one category could be strings which are totally dissimilar to each other, and then there could be a third category which could be things which are similar but could be distinct.

So interestingly, I think that still reduces the space for the panel because they can say, okay, things which are very similar to each other we can take a quick look and take it off and put it on the side that this is causing contention, and then strings which are totally dissimilar to each other like ABC versus XYZ, they can be put on this other side where we know that there is no contention and then let us look at those cases which are in the middle bucket in more detail. So, it can help expedite the process and that is the intention.

Moving on then next slide please. These are just some examples on how this may play out and we will go through this very quickly. So, we do not need to dwell into this. This is really up to the experts, and we will take this, the data we are gathering, we will bring it to public

comment so you will actually have a chance to comment on it whether you agree or not agree with the panels or the experts' opinion and based on that, experts will then finalize the data. So, these are just examples of category 1 from different scripts.

Let us move on to the next slide. Example for category 2 from Ethiopic and Latin. These are examples for category 3. Again, these are examples formed by data we are getting from experts but at a character level not at a string level. So please note that when it comes to string, the panel may actually differentiate or and it is eventually the panel's call. The tool is only assistive and it is only pre-screening and again, doing it at a character level. So, trying to add a sufficient amount of disclaimers here to make sure you understand that eventually this is really a manual process and under panels per view.

This is level 4 examples where I guess, they considered these to be not similar enough and there is a Hebrew example above. And again, we have to look at it as the guidelines say from a native script user's perspective, not perspective of the-- so, for example, I do not read Chinese, so if I am analysing strings in Chinese, to me my perspective is very different from a Chinese reader who understands the script. So, please go read the guidelines which we published which talk about some of these aspects.

And these are strings like ABC versus XYZ in Latin script, for example. So those are the five categories and as I said, currently we are doing the probability. The way we are translating the word probability is that we could divide it in 5 categories and 3 and above sort of provide a

probability versus 4 and 5 may provide a possibility but not quite a probability.

So let me stop here and see Tapani or if you still have a comment or others if you want to comment before we move on to the tool and process. And please let me know if I'm running over time.

MICHAEL BAULAND: Yeah. Sarmad, you're running a bit over time, but we decided that this is an important topic and the rest of the time-- Anyway not that much time left to start a whole new discussion so we are giving this topic a bit more time so everybody can ask their questions and we can then do the other the topic next call. Do you want to run the queue Sarmad for this?

SARMAD HUSSAIN: Go ahead, please.

BILL JOURIS: Yeah. And this is Bill Jouris for the record. You keep referring to a panel of experts. Is there some kind of definition for what constitutes an expert that will be on this panel versus a random person off the street? Have we defined who's going to be a counted possible expert? Thank you.

SARMAD HUSSAIN: Right. I think the process of setting up panels is through the regular ICANN process where there are-- And I think that is probably defined through the next round of New gTLD Program. And once those calls for panels are made, then at that time we will make that determination. Of course, experts mean that they have had experience in not just using the script but obviously analyzing the script the way we are analyzing. So, it's hard for me to give like a proper definition of an expert on the fly here but I guess we'll have to see the proposals eventually which come in and then make that call. Satish, please.

SATIH BABU: Thanks, Sarmad. Satish here. I apologize if this is a trivial question, but you mentioned underlined characters in the Unicode character set. Now what distinguishes underlined meaning formatting underlining versus scripts that have underlined built-in? Is that something that's confusable for end users? Thanks.

SARMAD HUSSAIN: So, what we were talking about was just the pure underlining process which happens to all strings which are treated as URLs in many different applications. We understand that now the applications are becoming even more advanced, meaning that they will actually break the underline if there's a diacritic at the bottom and so that diacritic is not overwritten but that behavior can change from application to application.

Again, idea eventually here is that the tool and the data presents all the relevant cases to the panel and then eventually the panel has to make

the decision. So, I think if we have to decide between being less conservative to versus more conservative idea for the tool and data would be to err on the more conservative side so that more of these examples are in front of the panel and then eventually panel can decide what to include in contention and what to omit. And I guess that's sort of where we're coming from. Thank you.

So, let's keep moving on because we are short on time here. So next slide please. These are the list of scripts for which the data is being prepared. These are the 26 scripts we talked about, and you'll see that data come out for public comment so please do look carefully at the data. This is for the community to agree with and so eventually we will use the community input to finalize.

So, this is then the process. Once the new round is applied for, TAMS is the application system where an applicant will come and submit their strings. Once all the strings are submitted, they are sent to this String Similarity Evaluation tool. We call it the SSE tool. And the SSE tool will basically take the data from the experts which is the SSE data. It will take the applied for strings which are coming from TAMS or the application system, and it will take a list of all the top-level domains, block names, all these other categories which are in some ways identified in the policy and it will crunch the applied for strings against each other as well as all these TLDs and block names using the SSE data to list basically possible contentions between strings.

And it will do it at two different ways. One, it will list all the contentions which have only characters which are level one-- between two strings which only are composed of characters which are either level 1, level 2

or level 3 different and not level 4 or 5 different because then there is a likelihood that those two strings are actually similar.

But in addition to that, it will also provide a string level measure where it will calculate a threshold over the entire string using the scores. So even if the two strings which are-- if you have a similarity score, let's say there's a three-letter string and similarity score is two and two between two strings on for the three characters level 2, level 2, level 2, that will automatically go, for example, into a potential similar those two things will be potentially similar. But if there is something which is, for example, one, one and four, that could potentially also be similar. And if we just say there's a threshold or that the similarity has to be cut off at three, those would not be available to the panel to review.

So, what we are saying is then that we will actually calculate a threshold in addition to the individual character scores, a threshold over the string which is an average value of those scores. And that threshold will determine another layer in addition to the character level similarity layer. And then this panel manager in red, they can actually generate a shortlist of similarity strings based on different thresholds. So, they can set threshold at 0.5 or 0.7 or 0.3. They check different thresholds and they can actually do different threshold for different scripts and so on. So that this tool will allow them to experiment not on character level only but also at threshold level, so string level. And based on that, they will actually generate first and initial shortlist or pre-screening.

So, the first loop here, back to the tool, is allowing them to play around with thresholds to see what is a good threshold where they think most of the data which or the similarity data plus a bit more is there, so they

slightly over generate but try to not over generate too much. And then in the second phase where you see that second red panel member on the right, then they look at that shortlisted and do a quick detailed manual pass and confirm the final contention sets, maybe take out a few and create the final report. So that's sort of the process. Panel members, of course, engaging with the tool and then using their own manual interpretation to make the final decision. And let me stop here to see if you have any questions.

MICHAEL BAULAND: There was one question in chat, Sarmad, asked whether there was any form of an appeal process to the decision of the panel.

SARMAD HUSSAIN: I think there is. Actually, I'll have to go and confirm. I think there are generally challenge mechanisms. And again, Saewon wrote in the chat that Ariel answered that there will be challenging mechanism of string similarity evaluation as well. So, thank you, Ariel and thank you, Saewon. Let's move on.

MICHAEL BAULAND: We just have a few minutes left. I don't know how much more you plan to--

SARMAD HUSSAIN: Let me take a minute and close. Thank you. So, this tool is being developed and we'll pass it all the needed data you see how it works.

This tool will be available for panelists currently and that's where we are currently focused on. And the tool will generate a contention set report as an output. It will be something which will then be integrated back into the TAMS, the application system.

I think this is just more detail which I've already covered. Next slide please. And this is sort of the output details of the tool. It will give contention sets for application 1, label 1, label 2 and the exact variant of that those two labels which are actually causing the contention. These are just examples of how this may look like. Again, we don't want to go into this. This is just process and evaluation results which will be processed by the application system. From a timeline-wise, we are very close to finalizing the data. We hope to have that available for public comment in May or latest by June and we are also now almost finalizing the tool and starting its testing and hope to close this work towards the end of third quarter of this year. Thank you. Back to you.

MICHAEL BAULAND: Okay. Thanks, Sarmad, for this very interesting and helpful presentation. Unfortunately, we're a bit of running out of time but the full presentation is also available. The link has been posted in the chat, and you can take a look at that offline too if you're having more questions. Also please note that this whole process is not for discussion here in this PDP. This is also just a given fact we have to work with so there's nothing we can change in that process at least not in this PDP. So, in case you're wondering whether this is how you think it should be and you're not happy with it you have to use other channels to change

anything in that process. So, this is just FYI for us. So, with this, I'll hand over to Saewon for the next step.

SAEWON LEE: Thank you, Michael. So, for the next steps, quite simple this week. I'd like to just remind everyone again that the early input request is by tomorrow. There is no extension. Another reminder was sent out at the beginning of the week to the leaders of each community group, and we have one response so far. So, looking forward to more by the end of this week. And then we didn't get to it today but in the next meeting we will be covering Charter Question 3, SubPro and EPDP IDNs. So please review the existing work especially focusing on EPDP IDNs Phase 1 and 2.

Again, same as last time that we met, another reminder that we won't be meeting on 7th of May due to the CP summit. When the early input comes in and again, we don't know how much will come in but based on how much comes in, the plan is to provide the working group with a high-level summary by or during the 14th of May meeting. And then again, just based on the project plan, we are proceeding as planned. So, I think everything is going as in order even with these weeks off. So, I think we're all good.

Again, this is the existing body of work again. You can find them all here or all the wiki pages that I've shared. And again, this is the project plan approved by the GNSO Council last week which is just for your reference. And now I'll hand over the floor over to John for the key outcomes and action items.

JOHN EMERY: Thanks so much, Saewon. So not too many outcomes or action items. Mostly a review of what we decided last time but important. So, Mark today shared his ASCII Unicode analysis report. This is not an authoritative source but it's helpful that we provide for convenience. We decided that allowing for multiple TLD versions for base ASCII requirements that we want to stay within the current scope, but we can recommend scenarios for future potential work in our recommendations.

So, an action item from that is later on in the process capture these cases that we determined to be out of scope, but we can recommend for future potential work. We all just thank you so much, Sarmad, for your presentation on the string similarity review process. I think it was really helpful for us to understand this. And then finally, we deferred Charter Question 3 until next meeting. So that's all we have for today. Michael, back to you.

MICHAEL BAULAND: Thanks, John, for the summary. So, unless there are any last questions and comments, I would also like to go--

MARK WILLIAM DATYSGELD: Michael, really quickly. Mark here. I would just like to point out that Ariel brought to chat the challenge mechanism. This is something that I would like to note in our meeting notes so that we have this available

for a future reference. This is something that's bound to come up again. So, thank you, Ariel.

MICHAEL BAULAND: Thanks, Ariel and thanks, Mark, for bringing this to our attention. And I would just like to thank Mark again for the presentation on his work on the Python tool and the tool itself of course. And a big thanks to Sarmad for his extensive presentation of what the string similarity review process will look like and how it comes to its data. And as mentioned before, this is nothing we can discuss here. It's just for our information. And that's it. And we'll talk again next week. Thanks, all.

MARK WILLIAM DATYSGELD: Thank you.

MICHAEL BAULAND: Thanks, all. Yeah, you can end the recording. Thanks.

[END OF TRANSCRIPTION]