ICANN Transcription

IDNs EPDP

Thursday, 11 August 2022 at 13:30 UTC

Note: Although the transcription is largely accurate, in some cases it is incomplete or inaccurate due to inaudible passages or transcription errors. It is posted as an aid to understanding the proceedings at the meeting, but should not be treated as an authoritative record. Attendance and recordings of the call are posted on agenda wiki page: https://community.icann.org/x/KgYVD

The recordings and transcriptions of the calls are posted on the GNSO Master Calendar page
http://gnso.icann.org/en/group-activities/calendar

DEVAN REED: Good morning, good afternoon, and good evening. Welcome to the IDNs EPDP call taking place on Thursday, 11 August 2022 at 13:00 UTC.

In the interest of time, there will be no roll call. Attendance will be taken by the Zoom Room. We do have apologies from Michael Bauland, Nigel Hickson, and Lianna Galstyan. All members and participants will be promoted for today's call. Members and participants, when using the chat, please select everyone in order for everyone to see the chat and so it is captured in the recording. Observers will remain as an attendee and will have view-only chat access.

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 e-

mail the GNSO secretariat. All documentation and information can be found on the IDNs EPDP wiki space. Recordings will be posted on the wiki space shortly after the end of the call.

Please remember to state your name before speaking for the transcript. As a reminder, those who take part in the ICANN multistakeholder process are to comply with the Expected Standards of Behavior. Thank you, and back over to Donna to begin.

DONNA AUSTIN: Thanks, Devan. Welcome, everybody, to today's IDNs EPDP call. I know I said at the end of last call that we will be continuing our discussion on—I just had it in my head and now it's gone. Continuing the discussion we were having last week on—

JUSTINE CHEW: Glossaries.

DONNA AUSTIN: No, not the glossary, the one after that.

JUSTINE CHEW: Oh, sorry.

DONNA AUSTIN: Strings in delegation. Thank you, Ariel. I had it in my head and then it's gone. But in a leadership call that we had, after the call,

we decided that it might be helpful to that conversation if we can develop some examples of how it would work. So we're going to do some more work on that before we come back to the conversation.

On the glossary, which was the first conversation we had last week, I just wanted to thank everybody for the conversation. I know from the leadership side, we had a conversation about what does this all mean and what's next. So we will continue working on the glossary. But in terms of the conversation we had around the primary TLD and what we mean by that, none of these settings are set in stone. Where we really need to be articulate is in the reports that we draft along the way. But I think we've come to agree that when we talk about the primary TLD, we're talking about the source TLD or the source label, which is we need a source label to be able to calculate the variants that an applicant could apply for, so the source label or the applied-for label.

I don't see any value in having a strict "you can't use primary again" arrangement. I know that with what Ariel and Justine are going to present to us today on string similarity, they still reference primary because that's the terminology we're using at the time. So I think what was helpful in the conversation is that I think we've reached a common understanding that when we talk about the primary label, we're talking about the source label or the applied-for IDN TLD label. To some extent, when we have the conversations on this group, those terms can be used interchangeably. But I think if we can try to head towards using source or applied-for label, it might be helpful for us when we're drafting. And it also goes away from the concerns that Dennis had

raised on the list about giving primary some kind of elevated importance when I don't think that's actually what we were trying to do.

On what we mean by set I think is a little bit more challenging for us. We'll have to be quite deliberate to explain when we say set, which we could be using to explain the applied-for label and the applied-for variants, we could consider that to be a set. But I think in the strictest sense, set is the string and the allocatable variants and the blocked variants for labels. So I think in the strict sense, set means a particular thing. But when we're trying to describe other types of sets, we will need to be a little bit more articulate about that.

So I think it's a good conversation. We will continue to work on a glossary. But I think if we can try to move towards using source label instead of primary, I think that would be good. But if you're like me and you continue to use the terminology of primary, then I think at least we all understand what we're talking about. So thanks for that conversation last week. I think it was really helpful. It's kind of evolved, we still got a little bit of work to do to hone in some of those terms so that we are very clear in draft reports that we put out for public comment. So any comments on that before we move into the substance for today's call? Ariel?

ARIEL LIANG: Thanks, Donna. It's not a comment about the glossary but administrative matter. It's about the meeting on 25th of August. Yeah. It will overlap with the GNSO Council meeting, and I believe some of the members on the call wouldn't be available, including

EN

staff. So we're wondering whether the team is okay, as in our previous practice, pushing the meeting 24 hours later, so we can reschedule to Friday, 26th of August at to 13:30 UTC. I just want to raise this quick matter.

DONNA AUSTIN: Thanks, Ariel. So we'll put that into chat, and then we'll come back to it at the end of the call and see what folks think about that. We'll also put it on the list. We've got a couple of weeks but it would be good to make a decision early.

Okay. So with that, Ariel, I'll hand it over to you. So the String Similarity Small Group has been working for six or eight weeks now. So it's been quite an effort and I think all of those that have been involved in it. Eleven weeks? Oh my goodness. So thanks, everybody, for the dedication. Obviously, it was a very difficult task. The group is going to present their outcomes here today. And apologies, I'm handing this over to Justine, not to Ariel. So, Justine, I'll hand it over to you.

JUSTINE CHEW: Thank you, Donna. I think Ariel could have easily and equally done a good job in explaining this. But since she wants to defer to me, that's fine. But I will ask Ariel to just draw my attention if she wants to add anything that she thinks I may have missed out when I'm explaining this. I would also like to ask everyone's indulgence, if you had a question or a comment, can you just make note of the slide number that pertains to your question or comment? But hold that until we're done with the entire deck, because you might find

that your questions are answered along the way. Okay. If I can just ask for indulgence on that front. Okay. The agenda is, as you see, we were tasked to do three things. We can just go to the next slide, I think. Next slide, please.

Okay. So this question that the small group handled came up some months ago and it's related to Charter questions E3, E1, and E3a, which is to do with essentially the string similarity review, what should happen with a review where variants are concerned, because now we're introducing variants of the [C2] TLDs. So in essence, things like what adjustments do we need to make to the string similarity review to take into consideration variants? What role, if any, should withheld-same-entity variants play? And what are the potential consequences for other allocatable variants in the same set of the requested variant label, which is actually rejected as a result of the review. Okay. So here it's talking about the set. Donna mentioned set in this context of the Charter question E3a. We're talking about the set of requested variant label. It would exclude any non-requested labels per se.

Then we also had a reference to the staff topic recommendation, which talks about comparing strings that are under consideration, not just against allocated or applied-for strings, but also the variants of those strings including allocatable, withheld-same-entity, and blocked. What this rolls out to be, we will look at it by way of illustrations and it will become clearer further down the track.

The EPDP team tried to tackle this question by way of looking at three possible levels of comparison. The focus was on visually confusable strings or visual confusability would be the main focus,

EN

and also tried to analyze impact and potential consequences. We derived three levels. Level 1 being primary and only requested allocatable variants. Donna mentioned earlier about the use of primary. In this context, we are talking about the source label, primary label being the source label or the label that is actually applied-for. That being the label by which the variant set is then generated or the other variants that's generated through the RZ-LGR for that particular primary label. So we're using primary in that context.

Level 2 consists of primary and all allocatable variants. So you can see the difference between Level 1, Level 2. And then Level 3 is the source gTLD and all valid variants. So that would include blocked as well as allocatable. Next slide, please.

By way of diagram, we will look at this very beautiful chart again that was done by Ariel. This is what illustratively means when we're talking about Level 1, Level 2, and Level 3. I'm not going to go to the specifics of it because I think we've been through this at least twice. But suffice to say that Level 1 is the least complicated and Level 3 is the most complicated because it could involve comparing many, many, many, many, many labels.

Okay. So what happened was then it was decided that a small team or small group should be constituted to actually look at the problem more precisely and try to come up with some solutions or proposed recommendations. So that's what happened. The String Similarity Small Group was set up. The two problem statements that were delegated to this small group was that there had been divergence of opinions regarding which level was most appropriate. So you recall some people favored Level 1, some

people favored Level 2, and I think possibly one or two people maybe favored Level 3. So there was obviously a divergence in opinion that way.

The other problem that was mentioned when we were discussing the EPDP level several months ago was that the discussion was largely based on academic basis and on abstract concepts. We needed some examples to actually examine what is it that we're talking about and what is the actual impact of what we're talking about? Next slide, please.

Then when the small group was constituted, obviously we made a call for volunteers and we had a number of volunteers. Then the first thing that we did was we came to an agreement on what we call the agreement sheet or the agreement form. Not an agreement, sorry. What's the word I'm looking for? Assignment form. In the assignment form, there were specified three tasks, which you see on the screen now. So the tasks basically involve developing concrete examples of variants, which would be considered as visually confusable, in order to better facilitate comprehensible discussion at EPDP level, which is what we're trying to do here today.

So the three tasks that were identified was—the first one is to develop the concrete examples. The examples had to be practical and pretty much limited to visual similarity because we're talking about confusability. It needed to be realistic enough that it could happen and get a sense of what could be the conclusions or the outcome if such case were to happen. We also noted that if anything were to fall through, then there could be existing

mechanisms like objections that could help catch other things that may have fallen through the cracks, so to speak.

The second task was to actually demonstrate how these examples would be dealt with in string similarity review in accordance with the three levels that we showed earlier, and then showcasing the impact of each review of each label that we selected and the potential consequences. Again, the view was we had a narrow remit. We looked from the aspect of security and stability and user confusability. So that was the narrow remit that we tried to keep to, which means that we did not look into or considered the complexity in implementing this possible solution that we're proposing today. Because number one, that's wasn't in our assignment form, and number two, I personally thought that that kind of discussion would be better to have in the overall, the full EPDP team than at the small group level.

The third task is to obviously see whether there was any differences when you apply the examples through the objection process. The Task 3 is something that we're going to be discussing at a later call. So we're not going to cover Task 3 at the moment because this small group had just concluded its work yesterday on Task 3. So we need some time to just clean up the recommendations in the deck. So we're just going to concentrate on Task 1 and Task 2 or mainly Task 2, really. So moving on to the next slide, please.

This is just to show you who actually volunteered for the small group and the language proficiency. I'm glad to say that we had quite a few languages that were covered, not just by way of script but also spoken language. So you have to remember that one

script could cover a few spoken languages. So, for example, the Arabic script could be applicable to Arabics because Urdu speakers and—I forget the rest but I'm sure Sarmad can point us in the right direction. But the point being that one particular script can cover a number of languages. And then even with one script like, for example, the Chinese, you have the differentiation between traditional and simplified Chinese.

We have had a total of 11 meetings. We started off a little bit slow but we were trying to make sure that we could develop a reasonably good way of showcasing what we're all about and what we're trying to get to and the results of it. So we ended up having these 11 meetings. Including yesterday was 11, right. Of course, apart from the members who had various language proficiency, we were also supplemented by staff who had also the language expertise like Pitinan and Sarmad and also Ariel. Okay, next slide, please. All right. So let's move on to Task 1. Next slide, please.

You remember the Task 1 was to find some examples. So that's what the small group did. We basically came up with eight examples. The examples, I believe, were mentioned during ICANN74 when we gave an update. So we actually reviewed this list of example labels that the small group had come up with. I'm going to ask Ariel if she can just speak to you guys a little bit about the practicality consideration column.

ARIEL LIANG: Thanks, Justine. Also, I just want to note that the green highlighted example showcase that it's, I guess, a cross script or

cross language for these examples. So the first example, the biss one is the Latin script one for Label A [part four], Label B as Cyrillic. And then for the example eight, the Label A is Japanese and the Label B is Chinese, but they're actually both using the Han script. So they're not extremely cross script but at least cross language. So that's why we want to highlight that.

In terms of practicality consideration, maybe I can just go one by one. The group really worked hard to try to find examples that could potentially happen. So for the first example, there are valid strings based on RZ-LGR calculation, but we cannot say for sure there are definitely real words in, for example, German or Cyrillic, these scripts for language. But at least they're valid strings based on RZ-LGR calculation. This is one of the first examples the group worked on, so we started at the beginning.

Then the second example, it's the HSBC in Chinese. Label A is the traditional Chinese version, and then Label B is simplified Chinese. So they're definitely real Chinese words, real brands. Then the problem with this example is that they're both variants to each other. So it's not really useful for our deliberation. So we have this example but we couldn't really use it for string similarity review-related deliberation. So that's example two.

Then for example three, I don't remember what this is in Arabic, but Sarmad can probably type in the chat if he remembers. This is actually proposed by Sarmad, I believe. So these are definitely valid strings based on RZ-LGR calculation, and at least one of these two labels has some meaning in Arabic script language. So it's definitely valid and they're not variants for each other. They are different words. So that's example three.

Then for example four, that's a Chinese example. The first one is simplified. It's, I believe, an artist's name, actually. Then the second one, Label B, is a trademark. It seems traditional Chinese. And they're definitely real words with different meanings. They're not variants for each other.

For the fifth example, they're very similar to the first example. They're basically Latin script examples and they are valid string space on RZ-LGR but we cannot say for sure whether they are real words in whatever Latin language it is. So that's why we didn't go further with this example so the group considered it.

Then the sixth example is Arabic example. I believe it's throw that would mean for Label A. Label B, I'm not sure what it is, but maybe Sarmad can provide some information on that. But based on RZ-LGR, they're definitely valid strings, and at least one up the strings has meaning in an Arabic script language. So that's example six.

Then example seven, it's kind of similar to example four. So the first label is Huawei, which is the Chinese brand, the telecom brand. That's definitely a real trademark. Then Label B here is basically repeating example four. It's an artist's name. Then Label C is the simplified version of Label B in example four. So if you just look at them, maybe as a non-native speaker, they already look confusingly similar, but there are definitely different strings and with real meanings and very different from each other. So that's example seven.

Then example eight. So the first one is EQ. It's a Japanese monk, that's his name. Then Label B is the traditional version of one

unity. That's what it means in Chinese. So there are definitely real words and in different languages and with different meanings. So that's the practicality concern when we look through the examples. Then in the following slides, Justine can provide our explanation of which example the group use to further deliberate on the string similarity review-related questions. I will stop here.

JUSTINE CHEW: Okay. Thank you, Ariel. A couple of things. One is I think in terms of these two columns of Label A, Label B, and the one Label C, some of them may not be variants of each other. But there may be elements of confusability when you start introducing the variants. So that's a simple answer to what Dennis was trying to, I think, get in the chat. So we're not looking actually for two labels that are absolutely variants of each other but labels that could possibly cause confusion. When we get to the actual examples, I think it will be clearer. What Ariel was saying is we ended up just working on example six and seven in order to do the demonstration of what we are proposing.

So, just to let the cat out of the bag, the small group is proposing something called the hybrid model. The demonstration of the hybrid model will come when we show the example six, which is the next one anyway, and then we also look at the possibility of three labels and applied-for label doesn't have to be just compared with one, it could be a number of applied-for labels that could end up being in contention set or something along those lines. It could be three or more, but we just kept it to three so that it remained manageable when we were trying to show the things what happens.

We also looked at the example of an existing gTLD versus an applied-for gTLD, and also to applied-for gTLD. So one could be existing and one could be not and the combination of different scenarios, I suppose.

All right. So let's go to the example six, which is the Arabic example. So in this Arabic, what we did was we ran the two labels through the RZ-LGR—Pitinan did. We came up with these two sets. So when we talk about sets, we're talking about the full source plus allocatable plus blocked. All right. So we came up with two sets of labels. I'm going to start using primary here again. So we have primary A1 which has allocatable labels of A2 and A3, and A4 to A24 of blocked variants. With the example of primary B1 which has no allocatable variants, but B2 and B32, blocked variants.

So what happens with the so-called hybrid model is in words we're just comparing primary with primary, primary with allocatable, primary with blocked and vice versa, so the other way around as well. But we would not compare blocked with blocked. So I'm going to ask Ariel to run through the actual steps of how we do the comparison. Can you do that, please, Ariel?

ARIEL LIANG: Yes, of course. So the actual step looks like this. You probably want to look at this graphic from left to right. So the first comparison is primary A1 is compared against primary B1. You probably recall B1 doesn't have allocatable variants, it has blocked variants. So then next one is for primary A1 compared

EN

against all of the blocked variants of B1, which is B2 to B32. So that's the first comparison.

The second comparison is compare the allocatable variants of A1 first against primary B1, and then allocatable variants of A1, which is A2 and A3 compared against the blocked variants of B1. So that's the second comparison in the middle.

Then lastly, the third comparison is to compare the blocked variants of A1, which is A4 to A24 only against the primary B1. So that's how it will look like with regard to the example that we just saw. It's with the set of variants that's under consideration.

So if you look at this consolidated view, that's what the arrows look like. So primary against primary, primary against blocked, then allocatable against primary, allocatable against blocked, and then blocked against primary. So the only part that's not being compared against each other, again, is the blocked against blocked. So that how it looks like. I guess I can probably cover the rest for this.

You see some of these arrows, they have this red color, and also the yellow highlight shows the number related to the comparison. So what it means is that during these comparisons, some visually confusable strings may be found. So for example, if you're comparing primary A1 against the blocked variants of B1, you may find that A1, B3, and B6, they look confusingly similar. I won't talk about every single one here, but on this little box on the right on this slide showcase possible pairs of confusingly similar string based on this hybrid model of comparison. I guess so over to you, Justine.

JUSTINE CHEW: Thanks, Ariel. So we're not saying that this is exactly what is going to happen. We're just saying potentially this is what might happen. And of course, it's up to the string similarity review panel to actually determine what is considered as confusingly similar strings. We're just using the example here to elaborate the potential consequences of leaving out a certain level or including a certain level.

Dennis, you have your hand up. Is there a particular point that you want to make? I wonder if we can just finish this slide first.

DENNIS TAN TANAKA: Oh yeah. Yes, sure. I can wait. Thank you.

JUSTINE CHEW: Thank you. All right. So the point being here is if we use this hybrid model, potentially four classes of confusingly similar strings could be identified. It could end up being that as the potential outcome of this says that the set of A1 and the set of B1 could get processed in a contention set. So using this hybrid model, that is potentially what could happen. If the hybrid model was not used and the blocked variants were not taken into consideration, so meaning if we say we just stick to Level 1 or strictly Level 2 and disregard those blocked variants, the ones in the pink boxes, then there could be potential of misconnection risk. Because in that situation, if you don't look at the two pink boxes, you just look at the A1, B1, and A2, A3, then potentially A1 and B could both be delegated. But they could lead to misconnection risks when you

look at the variant down the line, in such a way that a user may mistake A1 with B3. So because if B1 is delegated and even though B3 is a blocked variant of B1, the point is that the person who is looking at it may not know that, wouldn't know that it's meant to be blocked and may mistake it to be something else, which is why it leads to misconnection risk. Misconnection, meaning that they think it's something but it's actually something else, and they get sent to a site that is controlled by a different party to what they were expecting. So I might pause here and ask Dennis to maybe pose this question or make a comment.

DENNIS TAN TANAKA: Thank you, Yeşim. It took me a second to find the unmute button here. Yeah, I was writing on the chat, but I think this is faster. So what we're trying to solve here is to find the basis for comparison for string similarity review, which is only focused on visual identity, if you will.

I have a question as to what's the value of comparing against any string to a blocked variant label, knowing that a blocked variant label per the RZ-LGR is meant to not be delegated, not for delegation. So it shouldn't be applied-for at the beginning because the RZ-LGR he will say it's invalid. Let's say we do make the comparison that there are two possible outcomes. So one is the applied-for label, that the desired label is confusingly similar to a blocked label or not. So two possible outcomes. In the case of it's not confusingly similar, well, everybody's happy, right? There is no contention there. But what's the next step if it's found to be confusingly similar to a blocked label, which is, by definition, is not allowed to be delegated? Does it matter? That's, I guess, the

EN

question that I have in my mind. What's the next step? If we are considering making a comparison to a blocked label, what's the next step if it indeed is found to be confusingly similar? That's the question, Justine. Thank you.

JUSTINE CHEW: Okay. Let me try and answer that, and Ariel can supplement if she finds the need. The first question is regarding a blocked variant. You mentioned a blocked variant is not being capable of being applied-for or being delegated. We understand that but we were looking at it from a point of view of the end user, who, as I said, wouldn't know whether something is blocked or not. But they still mistake something that is delegated for something that they think is something else and therefore there is still a misconnection risk, whether the label that they thought it was is actually blocked or not. I hope I'm being clear. It's not a question of the fact that the label is blocked and therefore cannot be delegated. It's a question of the end user mistaking that label. They don't know two hoots about or they don't give two hoots about whether it's delegated or not delegated or blocked. But they think that label is what it is. They mistake it for something else and they get redirected to something else, which is what we're saying is the misconnection risk.

Now, the second question that you raise is what happens? It depends on what the result of the review is. It could be that something is not delegated or something, or the two sets get put into contention set and they get resolved that way. I don't have a definitive answer for you but those are the two possibilities that I

think the small group considered. Ariel, would you like to add something? Please go ahead.

ARIEL LIANG: Thanks, Justine. I think the outcome is very much consistent with what string similarity review outcome will be. So if it's two applied-for strings that are compared against each other and a confusing similarity is found, then they're putting the contention set and try to resolve in that manner. And then if it's applied-for a string and an existing string compared against each other, and then confusing similarity as found based on this hybrid model, then it's possible that the applied-for string is ineligible to proceed to the next step. So it's basically that outcome. We probably also need to consider SubPro's recommendation regarding the challenge mechanism, and it may impact the timeline and how quickly the outcome will become reality. So we also need to consider that. But it's basically consistent with the practice right now.

JUSTINE CHEW: Thank you, Ariel. This is what the small group landed on, really. Obviously, it's not set in stone. This is something that we're bringing back to the EPDP team as a whole, and that's what we're discussing. But we were tasked with finding examples and coming up with the possibility of what to do in terms of string similarity review with those examples. So this is what we've done.

I wonder if we can move on to example seven. Ariel, could you take us through example seven, please, if you don't mind.

EN

ARIEL LIANG:	Yes, of course. I will try to enlarge this. Example seven, we will use a different way to present it. It's a vertical way to look at the comparison. So it's these two Chinese strings. A1 is the trademark of the artist. It's [inaudible], and then B1 is another trademark, [inaudible]. Now we also want to demonstrate this particular scenario is that B1 is an existing string and A1 is applied-for string. B1 has allocatable variants and blocked variants, and A1 also has allocatable and blocked. So when you look at the comparison, it's basically A1 (the primary) is compared against the B1 (the primary), B2 (the allocatable variant of B1), and then B3 to B12. They are the blocked variants of B1. That's the first comparison. Then the second one is the allocatable variant of A1, which is A2 as compared against the B1 (the primary), B2 (the allocatable variant), and then B3 to B12, the blocked variants of B1. That's the second set of comparison. Then finally, A3 to A6. These are the blocked variants of A1. They're compared against B1 (the primary) and B2 (the allocatable variant of B1).

Again, if you look at the consolidated view, everything is being compared against each other. But the only set that they're not being compared against each other is that blocked variants against blocked. So that's the consolidated view. It looks very much similar to how the hybrid model looks like for the Arabic example two. It's basically the same logic.

In the following, you may find that when you look at the comparison, some of the strings that may be confusingly similar will be detected. This example, unlike the Arabic example, it's not the most perfect one, because you can already see confusing similarity against primary against primary, and allocatable against

EN

primary. For example, you don't need to go all the way down to blocked to find the confusing similarity. But still, we want to just demonstrate based on these level of comparisons, you may find these pairs of strings that may be confusingly similar.

Just to answer, I guess, Dennis's question earlier, what's the potential outcome if applied-for string and existing string are being compared against each other in the string similarity review? So due to the finding of these confusingly similar strings, one potential outcome is that A1, the applied-for string may be rejected or ineligible to proceed because it's found confusingly similar to an already delegated string, which is B1. This is the first scenario regarding this Chinese example.

Then the next one we want to demonstrate is that the scenario when we have three applied-for strings that may be confusingly similar, how the comparison will look like when using the hybrid model. So you could see that the things they need to compare it against each other will increase quite exponentially because we add another string into the, I guess, matrix or add to the comparisons. So just mathematically, the things I need to compare it against multiply.

Again, I want you to just quickly explain that first we have A1, B1, C1. These are the three applied-for strings. All of them have allocatable variants and all of them have blocked variants. So what you first compare it against is A1 (the primary) compared against B1 (the primary), B2 (the allocatable of B1), and B3 to B12, the blocked variants of B1. Then you compare A2, which is the allocatable variants of A1 against B1 (the primary), B2 (the allocatable), and then B3 to B12, the blocked variants of B1. Then

EN

you compare the blocked variants of A1 against B1 (the primary) and B2 (the allocatable). So that's just the first set of comparison when you do two strings.

But then we also have C1. So you repeat the same kind of hybrid model of comparison. So you use A1 (the primary) compared against C1 (the primary), and then C2 (the allocatable variants of C1), and then all of the blocked variants of C1. Now you do the allocatable variant of A1, which is A2 against C1, and then C2 and C3 to C9. Finally, you do the blocked variants of A1, which is A3 to A6 and compare that against C1 the primary and C2 the allocatable variant of C1.

Then finally, you have to do the same kind of comparison between C1 and all of their variant labels as well. I won't repeat how it looks like but it pretty much looks the same like the other two. So that's why you will potentially see 24 different comparisons in this hybrid model.

So if you look at the consolidated view, it looks like this. It's a little overwhelming to look at so that's why we colored the arrows and numbers in different colors so you can see what is being compared against the different segments of comparison, I guess. Then the only thing that's not being compared against each other is blocked against blocked. But everything else is being compared against each other.

Again, using this hybrid model, you will find quite a few confusingly similar pairs of strings. This is, again, not the best example because you don't need to go all the way to blocked variants to find confusingly similar pairs. You could potentially find

confusing with similar pairs if you just compare the primary to primary or allocatable to primary. So that's possible to find that. But if you use the hybrid model, you could potentially find many more confusingly similar pairs where you consider the blocked variants.

Then for this potential outcome, because all three strings are being applied for during the same round, then they may get processed in a contention set, because the confusingly similar pairs of strings are found due to this comparison. That's the second scenario when we take into consideration three strings, and then of course the hybrid model will increase the things being compared against each other quite a lot. I will stop here.

JUSTINE CHEW: Thanks, Ariel. Based on those two examples that Ariel has very kindly explained to us or run through with us, this is the recommendation that the small group landed on in terms of string similarity review, is to use what we termed as a hybrid model, which is basically a mixed level approach between Level 2 and Level 3. The goal behind this hybrid model is basically looking at the mitigation of possible confusingly similar cases between one TLD and another TLD and its valid variants, valid variants meaning both allocatable and blocked.

Again, in text or word terms, this is what it ends up being, that in practice, we're suggesting that the string similarity review ought to be modified to compare the applied-for string label and all of its allocatable variant labels against all existing TLDs and their allocatable and blocked variant labels. Also strings requested as

IDN ccTLDs and their allocatable and blocked variant labels. I should be consistent with the term. Sorry. And also other applied-for gTLDs in the same round and all the allocatable and blocked variant labels. Also reserved names and any other two-character or two-char ASCII strings and their set of allocatable and blocked variant labels.

In addition, we also need to look at all blocked variant labels for an applied-for primary IDN gTLD against existing TLDs and all their allocatable variant labels. As Ariel and I have pointed out, the only two groups that we do not compare against each other would be the blocked variants.

Moving on to the rationale behind the hybrid model, if it's not already apparent, again, we're looking at the limited narrow remit of security stability, user confusability. That was one of the things that the group was most concerned about in coming up with the hybrid model, which is the user confusability aspect. Because of that, the group believes that the hybrid model is sufficiently conservative and can help mitigate the two failure modes, which is, one, denial of service and the other one is the risk of misconnection, which we thought had the higher likelihood of affecting non-native speakers of certain scripts or languages. Again, it's a question of, say, for example, an Arabic label being misinterpreted by somebody who doesn't speak Arabic but speaks another language but uses the same script.

The hybrid model, we also believe will help detect more pairs of visually confusable strings, and therefore reduce the risk of failure modes. Again, denial service or risk of misconnection. Because we noted and we also thought that by comparing blocked variants

EN

with blocked variants, that would lead to exponential level of complexity when it comes to comparisons. I think, all of us, if not most of us, didn't think that we need to go to that level of complexity when it comes to blocked against blocked. But we thought that we would still include everything else in the comparison matrix in order to maintain the level of conservatism that we thought was sufficient, basically to reduce the risk of failure modes.

The additional considerations would be that the pool of strings will be considered by language experts in the String Similarity Review Panel. As before, we will be expecting them to weigh the strings on a case by case basis. We also noted that there were other mechanisms, as I mentioned before, such as the limited appeal and the objection processes. Limited appeal in terms of the evaluation of the String Similarity Review Panels result or decision. And also objections by parties who are impacted by any strings passing through the string similarity review as proceeding onwards in the process.

I think that brings us to the last slide which is again the repeat of the recommendations, I think it was. But I've already gone through that so I think that's fine. That's it. Okay. I will hand it back over to you, Donna, unless Ariel has got anything else to add.

DONNA AUSTIN: Thanks, Justine. Thanks to everybody on the small group that's worked on this. Obviously, a lot of thought and effort has gone into it. What I'd really like to do is open up now for questions, and also how do folks feel about this recommendation is coming from the

EN

small group. Because ultimately, what we're asking here is if it is a recommendation that we can adopt as the working group and move forward with this as our plan for the string similarity review. Justine?

JUSTINE CHEW: Thanks, Donna. Before we get to that, I just like to take the opportunity to thank all the small group members, as well as staff who provide an immense support, as usual. I would encourage members of the small group to speak to any point or try and answer any questions that the other members of the EPDP have. Thank you.

DONNA AUSTIN: Thanks, Justine. With that, if any members of the small team would like to add anything to the presentation that Justine and Ariel have just provided, please do so now. But we really want to hear from the rest of the working group. Hadia, go ahead.

HADIA ELMINIAWI: Thank you so much. Thank you, of course, to Justine for this presentation, and thank you for staff for all the help. I just wanted to note one thing, and I think Justine already did when she said that we thought that the implementation and how complex this could be when being implemented, we thought that this is something we need to leave for the full group to describe. And if we look here at this slide that we have in front of us, it says an applied-for primary IDN gTLD and all of its allocatable variant labels are compared against existing TLD strings requested as

EN

IDN ccTLDs and all of their allocatable and blocked variant labels. Adding the blocked variant labels here could actually lead to increasing the number of comparisons by 1000. There are specific examples. I think they were presented by the ccTLD group during the mutual meeting. This is the only thing I just wanted to point out. Again, I'm not leading towards any kind of decision. It's just that this is something we need to take into consideration as well. Thank you.

DONNA AUSTIN: Thanks, Hadia. A question I have. I'd really like to understand a little bit more about the misconnection risk. Because similarly to what Dennis raised, I'm not sure I fully understand the value of cross-referencing against blocked variants that aren't going to come into play anyway because they can't be allocated, they're blocked. But I'd like to understand more about the misconnection risk, if someone could explain that.

JUSTINE CHEW: Can I send this off to Sarmad? I think he has the best example by which to illustrate this. Sarmad, would you please? Thank you.

SARMAD HUSSAIN: Thank you. If you can go back to the slide with the Arabic example. Actually, I think the one before it. Yes. Let's look at B1 and its variants. So, assuming that it is a valid label RZ-LGR, assuming that somebody applies for it. What the blocked variants in pink by definition suggest is that they are considered somehow equivalent to B1, all the blocked variants B2 through B32. So,

somebody who's looking at B1 could, for example, consider B3 the same as B1. That's by definition of variants.

DONNA AUSTIN: Sarmad, can I just ask a question there? When you say "somebody," who do you mean?

SARMAD HUSSAIN: Actually, it is the end user who considers these two, I guess, characters as variants within the Arabic script against which the Arabic script community suggested that these two characters will be considered variants of each other. So, whenever two different code points are defined as variants of each other, either they are visually identical, in this case, they're not, or they are considered equivalent or "same" by the script community for one of the particular languages or some use of that script by a particular language community. So we are obviously talking about that particular language community against which these two letters which are causing these variants, let's say, B3 and B1 were considered variants by the Arabic script community. So it could be one language community, it could be more than one language community within the script community. To answer your question—

DONNA AUSTIN: Yeah. Sorry, go ahead.

SARMAD HUSSAIN: No, please go ahead.

DONNA AUSTIN: Have you finished with the explanation?

SARMAD HUSSAIN: No.

DONNA AUSTIN: Keep going.

SARMAD HUSSAIN: So what I was saying was that for some language community, let's say, B1 and B3 are considered variants, therefore, they are considered same, right? If you have a user of that language community, and let's say that for some reason we are not comparing blocked variants, it would mean that from a string similarity perspective, there will be—if B1 is applied and A1 is applied now, not looking at B3 anymore because that's what's being compared with, they will not be considered visually identical. And therefore, both A1 and B1 will proceed to delegation without any contention set. If that happens, then coming back to the previous scenario, which I was discussing, somebody from that language community which considered B1 and B3 to be the same, when they look at A1, they think that A1 is the same as B1 because they consider B1 and B3 the same. So we now have two TLDs which are delegated in the root zone, A1 and B1, which were some subset of users of Arabic script would consider A1 to

EN

be same as B1. That is why I think it was being suggested that the blocked variants should be taken into account. Otherwise, those users, I guess, could be put into harm's way because of security issues, because they could look at A1, they would think it's B1, and therefore, they'll click on it but they will go to some other place which is different from B1. That does not apply for one domain name but it applies for, for example, all the domain names which are at second level registered under the TLD A1. Thank you. I hope that clarifies.

DONNA AUSTIN: Thanks, Sarmad. I think I have a better understanding. But I want to get back to Dennis's note in the chat that misconnection happens every day when users type a different extension to the second level domain. It becomes obvious when you click on that link what the intent of the TLD is. I understand what you're saying but that could actually be happening now. I don't think it's unique to IDN labels. I can see that A1 and B3, it is visually similar so that could cause some confusion. That misconnection piece, I think I tend to agree with Dennis. I think it's happening now. So what's unique about the IDN labels that we want to mitigate against that? Justine?

JUSTINE CHEW: Thanks. Perhaps I could try and answer your question. Well, yes. It's true that misconnections happen every day, but we do not know whether it is as a result of mistyping, or whether it's phishing or something along those lines. Of course, I can't say what we can do about existing TLDs and the problems with that. But in terms of

what we can do with IDNs and variants is this little bit that we're trying to put forward, which is to help reduce the risk of misconnection. We thought about it in the frame of harm that could come towards the end user, not just inconsistency in terms of experience, meaning to say that they get led to somewhere that they didn't anticipate that they will get led to. And there's also the thing about phishing and that sort of thing, pharming, phishing, harm that we were thinking of. Thank you.

DONNA AUSTIN: Thanks, Justine. I know this is a lot to take in. This is the first time that most of you have seen this. I'd really love to hear from other members of the small group that were part of this effort to explain to us why you think this is a good idea. There's another explanation I think where I've gone to at the moment is what happens when you do end up with a confirmed string similarity, because I'm wondering here with the examples of B1 and A1, and why the meanings of those might not be the same, but I can see that B3 looks a lot like A1. So what happens when that's picked up? What's the next step? And how do we resolve that string similarity? I guess they go into contention set and they have to sort it out. I guess my thinking is that because of the number of blocked variants, this might happen potentially quite a lot. Justine?

JUSTINE CHEW: As I said, I would encourage all the other members of the small group to pitch in. But again, trying to answer your question, if A1 was an existing TLD, then the answer to your question would be that B1 would be rejected. If A1 and B1 were being applied for in

the same round, then they would probably both go into contention set.

DONNA AUSTIN: Okay. So anyone have any other thoughts on this? Ariel?

ARIEL LIANG: Thanks. I see Dennis has his hand up. But I just want to note one other thing that I was also mentioning the rationale is that, at the end of the day, these strings need to be evaluated by the String Similarity Review Panel that consists of language experts. So the examples that the group developed may look daunting. But when it's in actual practice, maybe the language experts, they look at them and just think there's zero possibility, it can be confusingly similar. Even from non-native speaker's point of view, maybe there's some confusing similarity there. But at the end of the day, it has to be evaluated by the language expert. So it may not go to that level of complexity that hybrid model uses. So I just want to note that that's something considered by the group when they put forward this hybrid model is that in actual practice, it may not be too crazy based on how it looks like in the slides.

DONNA AUSTIN: Thanks, Ariel. Dennis and then Satish.

DENNIS TAN TANAKA: Thank you, Donna. Yes, I appreciate the work that has been put into this effort, Justine and team. There's a lot to process there.

EN

The Registry Stakeholder Group members will take this in, process it, and come back with our opinions or observations. Thank you.

DONNA AUSTIN: Thanks very much, Dennis. Satish?

SATISH BABU: Thanks, Donna. Can we go back to the previous slide that we were looking at? I think a lot of work has been done and I'd like to congratulate the small group. Not this one, the one that Sarmad was talking about. Just now. No, the one before this. Yes. Okay.

I think I understand what is really going on here, that you're not comparing the blocked with blocked. In the last round of our discussions, I was actually in favor for bringing in blocked as well because it would future proof. In the case of Root Zone LGR coming up with a new version where a blocked point was actually promoted to allocatable to ensure that that doesn't happen. So just for future proof this whole thing, we had thought of blocked also to be considered.

But what I am not able to understand here right now is that now, in this case, even not visually distinct, there is no room for confusion so they're both delegated. But one of the blocked variants of B1 and B3 is visually similar to A1. And the misconnection scenario seems to say that the end user, instead of typing B3, types A1. But my question is—because B3 and B1 are different, so why should the end user type B3 and why not B1? Because if you type B3, immediately there's fail fast kind of error, which is the DNS

EN

refuses to resolve. So as an end user, what might happen is I will go back to B1, in which case, I'll get the right site. But the assumption here seems to be that the end user, instead of typing B3, will type A1 because of visual similarity.

So to me, that is not entirely intuitive. Also, I didn't really understand the denial of service, which presumably the same thing. My observation basically is that if somebody types B3, immediately there is a response that says this is not a valid domain, and therefore, the end user has to reconsider what has happened. Thank you.

DONNA AUSTIN: Thanks, Satish. Sarmad?

SARMAD HUSSAIN: Thank you, Donna. Just to respond to Satish's concern. At least the way I was looking at it was not from a typing perspective but, for example, somebody using or sending a link to someone to click and they use A1. If you have example.b1 and I want to do some malicious activity and I know that there's this whole population of people who think example.a1 could be confusingly similar to example.b1. So I would register example.a1 and send the link to that community, and they would read it and they will say that "Okay, this is visually distinct, but it is the same." So they could potentially click on example.a1, thinking they want to go to example.b1, but the link takes them to example.a1 which is a malicious site. Thank you.

SATISH BABU:	Thanks. That makes sense to me.

DONNA AUSTIN:	Thanks, Sarmad. Okay. So we recognize there's a lot to unpack here and we want to give folks some time to go back to their respective groups and have some discussions around this. So I'm thinking, can we come back to this in about two weeks' time? Is that okay for folks? So it would be the meeting that we've pushed out the extra 24 hours. So the meeting is the 26th, we'll come back to this.

Okay. As you're working through this with your respective groups—I don't want to put Ariel on the spot—but perhaps if you have any questions, just put them to the list and we can ensure that we get an answer to help you through it all. I know Ariel is pretty receptive to helping people out. So if you need some more assistance in working through the tables, I'm sure Ariel would be more than happy to help out where she can. But of course, the recording of this will be available after we have this call, so folks can go back and have a look at that.

I think from the perspective of what's the most conservative approach and the one that will be perhaps the best in terms of trying to address not only confusingly similar issues but also some of the other things that Sarmad has referenced such as phishing and nefarious activity, I think that's good, but at some point, we're going to have to shift our headspace into what is implementable. So what I would hate to see happen is that we have 100 applications for IDN gTLDs and 20% of those end up in string contention set.

So the other thing that we might be able to look at, depending on how this pans out, is I know that within the SubPro, one of the recommendations for contention sets that apply to a brand is that there is an opportunity to change the string. So I wonder—and I am getting a little bit ahead of myself here—but perhaps that's something else that we could think of as a way to try to overcome that potential downside of this is that 20% of our applicants end up in contention sets. So is there an opportunity to offer them to choose another string to pull them out of a contention set? Would that work? So that's me thinking probably four or five steps ahead of where we are today. But I just wanted to put that out there because I'm sure that you'll all start to move into that implementation space and how we can manage that. But it's just to see that there are some other options that might be available to us.

So again, thanks to the work of the small team and to Justine for your leadership. It certainly was a heavy lift, and we can all see the good work that's come out of it, so we're very thankful for that. So we'll leave it there. I think I've seen a couple of folks in chat over moving our call on the 25th to the 26th works. So we'll put that to the list and say that this is what we're going with in the absence of objections, and then we'll come back to this on that call, which will be the Friday.

All right. Thanks, everybody, and thanks for your patience and attention. We certainly have some fun on these calls. All right. So with that, we will talk to you all in a week. We can end the recording. Thanks, Devan.

EN

DEVAN REED: Thank you all for joining once again. This meeting is adjourned. I'll in the recording and disconnect all remaining lines.

[END OF TRANSCRIPTION]