ICANN70 | Prep Week – Internationalized Domain Name (IDN) Program Update
Thursday, March 11, 2021 – 16:30 to 17:30 EST

PITINAN KOOARMORNPATANA: Thank you. All right. Thank you and welcome, everyone, to the IDN program update session during the ICANN70 prep week. Also I would like to remind you, if you're not speaking, please mute your microphone. Thank you. All right. My name is Pitinan Kooarmornpatana and I will take you through the session today.

For the session today, we will have three main parts. First will be the updates from the IDN programs by me, and then the second part will be the updates from the communities. So, for this meeting, we have participants from four communities.

We have Panagiotis from Greek Generation Panel. We have Kim Kyongsok from Korean Generation Panel. We have Bill Jouris from the Latin Generation Panel. We have Yin May Oo from Mayanmar Generation Panel. So, they will give the updates of the progress of the Generation Panel for the root zone LGR project. Finally, hopefully, we have ten to 15 minutes for the Q&A. Let me dive into the first part.

So, for the IDN program objective—this is just to reiterate—the goal is we will enable the deployment of domain names in the local languages and script for global communities. Of course, this has to be done in a secure and stable manner. So, to achieve that goal, we have several projects to do that, both for the top-level and for the second-level domains. So, for the top-levels, we have root zone label generation

rules, or root zone LGR projects, which is where the communities come together and define how to use the particular script properly for the root zone.

Second projects, then, after we have the definition, then how to manage them. We have the Variant TLD Implementation Project. And then, the third part will be the IDN ccTLD fast-track process. Then, moving on, for the second-level, we have IDN implementation guidelines projects and reference LGRs, and also the LGR or IDN table review tools. So, each of these, I will go talk a little bit about a background for the one who is not familiar with the projects and then give some updates.

So, for the first one, the root zone label generation rules, or the root zone LGR. This is the procedure which was developed by the communities to be able to come up with the rules how to use each script properly in the root zone. For this, this is the structure set out by the procedures: have two-step panels. First will be the generation panel, which is formed by the script users. Each generation panel will be comprised of various parties from the multi-stakeholder model, many of them comprised of technical experts/linguistic experts. Also, some have the ccTLDs and also registries and registrars for that script.

So, the task for these generation panels is to define how to use the script to generate the label for the root zone properly, which have to define three things. First, what are the possible code points to use? Second is, what are the variants, if any? Which we will talk a little bit further. And also, what are the rules that are required?

Once the generation panel has finalized their work, it will be published for the public comment. And after finalized, after taking into account input from public comments, then it will be sent to the integration panels for integration. So, integration panels will make sure that all the scripts can be integrated well.

And then, if there is some further discussion, they might have some interaction between the generation panels and integration panels. Once the integration is finalized, it will be integrated into the root zone. So, the root zone LGR is being integrated incrementally. Right now, we are at the root zone LGR version four, which includes 18 scripts.

So, this is the status of each generation panel. The ones that have already been integrated will have this dark arrow on the top. As you can see, many of the other script panels are very near to finalizing their work. Of course, we have two more scripts, which are yet to form the panels. But in general, this project is fairly close to the final stage. Today, we will hear the updates from the generation panels.

Moving onto the next project, the IDN Variant TLDs. The “variant” is something that can be perceived as the same by the script users. However, the definition of the same differs across the communities and it can have two different implications. So, first, it can mean for the security issues. The second would be for the usability. So, let me give some examples.

On this slide, if you see the epic in this blue text, and also another one in the blue text. To us, to the human eye, it looks exactly the same. But behind this, in this four-digit code point is something computers

understand. Basically, to the computer, these two strings are actually two different names. For this, it can create some security issues for the users because users can understand that they want to go to some place but end up in another website. So it can cause phishing issues, for example.

So in this case, the relevant community—and in this case, they are Latin GP, Cyrillic GP—have to come together and figure out that each character, the E here and the E lookalike here, are variants, to prevent security issues for the users.

Then the second aspect of variants is the usability. For example, here for Chinese characters, you can see that the second characters are slightly different, and that's because in Chinese, they have two versions, the traditional ones and the simplified ones. So to the Chinese community, they will understand that these two words can mean the same thing, but it's just the different character used in the different countries.

So if you are a brand owner, you would like to be able to use these two names simultaneously, so your customer can reach your website. So in this case, the Chinese community has to come together and define that these two are variants, and this is for the usability purpose.

And for the Arabic as well, as this is some case similar, you can see that the last characters are slightly different, but for the Arabic users, this can be understood as the same across the country. So this is the work that the generation panel has to define.

So, what are the scripts that need to work on the variant codepoints? Basically, as you can see, only four of them don't have any variants. Most of the scripts have some aspect of variants to think about. So this is not a trivial issue. And the work of the generation panels are very important.

Moving on, once we have the definition of what are the variants, then it's come to the next question, how to manage them. So for the variant management, ICANN Org has developed a recommendation which finalized in early 2019 which later was adopted by the Board in March 2019. The Board requests that ccNSO and GNSO take into account this recommendation when they develop their respective policies.

So for the status today, both ccNSO and GNSO has the working group related to IDNs going on. For ccNSO, there's a PDP for selection and deselection of the IDN ccTLD strings which also take into account this recommendation for variants.

For GNSO, they have the draft final report of the SubPro, also taking into account some of the recommendations and also currently, the IDN EPDP is starting and is at the drafting charter stage which is also taking this recommendation into consideration.

For the next projects for the top level, ICANN also implementing the IDN ccTLD fast track process at the string evaluation part. So far, we have successfully evaluated 62 strings from 43 countries and territories, and some countries might have more than one based on how many of official scripts they have in the countries.

Then next, for the second level, IDN implementation guidelines is the guidelines that we develop—the community develop, and the aim is to minimize the risk of cybersquatting and also the consumer confusion. For this, the guideline applies for the registry that offers the IDN for the second level. This guideline has been revised from time to time and the latest version was the version four, it's been finalized and published in 2018.

Later in 2019, GNSO has made a request to the Board that it would like to have some more time to study the guidelines before putting it into implementation. So the status right now is the GNSO is organizing the operational track to review these guidelines.

Then moving on, for the reference second level LGR projects, this is in continuation of the second level. So if a registry would like to offer the IDN labels for the second level, then they will need to submit the IDN tables to ICANN and then we will review for the security and stability issues before it goes to IANA repository. This is something similar to the—it is also the label generation rules, but it is for the second level.

So we develop the reference IDN table so that the registry can use it as a baseline. They can directly adopt if they like, or they can use it as a baseline when they decide on IDN table.

Right now, we have published 42 reference IDN tables. Many of them are script-based, which means it can cover a lot more languages using the particular scripts, and we have also a number of language base as well. Currently, we have four more in the public comment right now. It's closing today, actually. So they are Arabic script, Hebrew script, Sinhala

script, and the Hebrew language. So if any of you are interested in those languages and would like to take a look, please take a look there and provide your comment.

Also, for the last project, this is together with the reference LGR. So when the registry would like to use the IDN table, it doesn't have to be exactly the same as the reference. It can deviate, but it has to make sure that the deviation doesn't create any security issues. So we provide a tool to be able to assist that. So the functionality of the tool is the registry or general user can upload the IDN tables and then can select the reference LGR they would like to compare with, and the tool will compare in a way that it can identify if there are any place that it should be reviewed for the security issues. So for this, we hope that it could reduce the time of the review and give more transparency and consistency to the process. The tool will be available in April, and it's coming up.

Okay, and that's a very brief update on the IDN programs. Let's see if there is anything in the questions now.

SARMAD HUSSAIN: Hello. There are no comments or questions in the chat right now. Thank you.

PITINAN KOOARMORNPATANA: Okay. Thank you. Then let me move on to the next section. I would like to invite generation panels to give their updates. So for the first one, for

the Greek panels, the floor is yours. And please speak slowly and close to the mic. Thank you.

PANAGIOTIS PAPASPILIOPOULOS: Thank you, Pitinan. Nice to see you again, and hello, Sarmad, hello, dear colleauges and friends, and the interpreters. My name is Panagiotis Papaspiliopoulos. I'm the chair of the Greek generation panel. It is a panel that deals with the Greek script. It is used in the Greek language, the language that is spoken in Greece and Cyprus.

Our panel was formulated almost five years ago. We had many things to settle in the beginning. Next slide, please. This is the short bullet points of what I'm going to say tonight. Next slide, please. You can see some information about the Greek script and the Greek language, and we are using the MSR version four. As I said, the Greek script is used for the Greek language which is the official language of Greece and Cyprus. Next slide, please.

For that reason, the panel consists of members from both countries. I am the chair, Mr. Segredakis is the vice chair. He's from the registry. The members are from many different stakeholders. We have members from governmental bodies, regulatory bodies, standardization bodies, linguistics, and from both countries. We have as well made it formal, because we were under the auspice of the Greek competent ministry of digital governance. Next slide, please.

So, as I said, the panel was formed about four and a half years ago, and we had to deal with many things. First of all, we had to make sure that

no other language uses the Greek script. We had done an investigation related to languages used by communities who used to stay and live in Greece but that were actually [dead] language if I may say that.

There is also another living language. This is the [inaudible] language. This is another language. There was an attempt to use Greek letters to make it a written language, but after discussing with a scholar of this language, the panel decided that there is no need to change any outcome that we already had or to formulate specific rules especially for that.

So after having done this investigation, we started to analyze, let's say, scripts related to Greek, like Latin, Cyrillic and Armenian. Greek is an old, ancient language, and the Greek civilization influenced all the neighboring population around Greece. So it was unavoidable that the Latin script, Cyrillic script and Armenian script had some letters that were taken from Greek or they're very much alike the Greek letters.

So we had to analyze these cases in order to determine any cross-script variants and similarity, or if I may say so, similarity in a degree that we should classify as cross-script variants. And of course, in Greek language, we use accents. For example, my name is Panagiotis. I have to put an accent in O. There are cases where an accent is needed to distinguish the meaning of the word, like Athena is the capital of Greece and Athena is the ancient goddess of wisdom, so the accent actually plays role. And this is why we have concluded that we have in-script variants. Let's say the basic form of a letter and form with the accent.

And we also have a final sigma, so we have actually two forms of sigma, one that can be used in the middle of a label, a word, or a final one. After having analyzed a lot all these years in detail, any possible case related to Latin, Cyrillic and Armenian scripts, in the end, we proposed several versions, actually, but in the last one, we had discussions and with the valuable help of Sarmad and Pitinan, we came to the conclusion that especially for the root zone, we can apply a mechanism to limit the allocatable variant labels to two, the applied for label and the base form, and that means that for my name, if this was a label, it could be Panagiotis with the accent on O, but also Panagiotis without any accent.

So we decided to apply this mechanism, and we have an exception in the case of final sigma. So the maximal number of allocatable variants per original label is four, if there is an accent and final sigma too. In order to cover the similarity cases and actually to support any work that might be doner in the future, in the last version of our report, we have also an annex indicating this kind of similarity cases just for any future reference.

So, we feel that we have done a lot of progress. It was not an easy job. To be honest, it could be done a little faster, but we believe that we used our time efficiently to conclude to a safe result and a safe proposal to the generation panel proposal that can be combined with the proposals of the Latin, the Cyrillic and the Armenian scripts.

I will not keep you long. This is the plan. And it's more or less the plan that Pitinan has already shown. We are confident that we will meet all

the milestones and we will be on time according to the schedule. Thank you very much for your time, and I will be available for any questions, remarks that any one of you might have. Thank you very much. Thank you, Pitinan.

PITINAN KOOARMORNPATANA: Thank you, Panagiotis. So let's move on to the next generation panel. So Professor Kim Kyongsok, the floor is yours. Thank you.

KIM KYONGSOK: Thank you. My name is Kim Kyongsok and I'm representing Korean generation panel. It is early in the morning in Korea. Nice to meet you. Thank you. I make a presentation about Korean LGR. Next slide, please.

Okay, here is agenda. Next, please. Korean LGR covers Korean script that is Hangul plus Hanta. Korean script usually means Hangul. However, in the context of Korean LGR, Korean script is a union of two: Hangul and Hanta.

Korean language has long history spanning more than 2000 years. The script Hangul was invented about 600 years ago. Hanta was used before Hangul was invented. Hanta is still used in Republic of Korea.

Korean language is mainly used in the Republic of Korea and the Democratic People's Republic of Korea. Also used by people outside of two Koreas. Next slide, please.

Here is the membership. I'll skip this. And here is progress summary. Korean LGR version 2.1, we had about 11,000 Hangul syllables and no

variant groups among Hangul syllables. And regarding Hanta, there are 4758 characters, and there are 37 variant groups prior to expansion. After expansion, we have 283 variant groups. So among Hangul syllables and Hanta characters, we have seven variant groups. And among Hanta characters, we had 289 variant groups after expansion and a repertoire of 4758 Hanta characters shown here. That is basically a union of two, one is KS X 1001, and the other is IICORE in ISO/IEC 10646. Next slide, please.

In 2018, we had the first public comments, and the major point was that whether or not to allow Hangul-Hanta mixed labels and some negative comments and some positive comments. Next slide, please.

Okay, see the last bullet. We discussed whether we allow Hangul-Hanta mixed labels. For Hangul-only labels and for Hanta-only labels, there was no problem. So KGP did confirm that there was general consensus to allow Hangul-only labels and Hanta-only labels, however, KGP decided not to allow Hangul-Hanta mixed labels. Next slide, please.

Then we made the four policies, version 2.0, 2.1, 2.2 and 2.3. 2.3 is the last one, and we are currently awaiting public comment. Next slide, please. Here, it is waiting for the IP feedback regarding version 2.3, but actually, the IP comment came back and the K-LGR version 2.3 is posted for the second round of public comment on March 2nd. And it'll be closed in April. Next slide, please.

Here is a brief history of KGP activities. It was organized in 2013, so almost eight years ago, and we published K-LGR version 0.1 in 2015. In 2016, Korean community formally formed generation panel for

developing root zone label generation rules. In December 2017, we published the version 1.0, and in 2018, from January to March, there was a public comment period for K-LGR version 1.0 and we had about 39 KGP meetings and several CJK, that is, Korea, China and Japan coordination meetings during ICANN public meetings. And also in the Republic of Korea, China, and Taiwan. Next slide, please.

Here is the plan, and as you can see, in March 2021, we have a second round of public comment for K-LGR 2.3, and if everything goes well, K-LGR will be integrated into a subsequent version of the root zone LGR. Next slide, please. Thank you.

PITINAN KOOARMORNPATANA: Thank you for this, Kim. Next, I'd like to invite Bill Jouris to give some update on the Latin generation panel.

BILL JOURIS: Good morning, good afternoon, good evening. I'm Bill Jouris. Our chair, Mirjana Tasic is fighting off COVID and pneumonia, so I will be giving the presentation today. Next slide, please. Most scripts are used for a single language. If you are doing Laotian, you use Laotian script. If you're doing Georgian, you use Georgian script. There are a few cases where neighboring languages may share a script—Arabic and Farsi for example. And a couple of extreme cases, the Cyrillic script in Eastern Europe and [inaudible] script in South Asia where you have a half dozen languages in the same area using the same script.

Then you have the Latin script. 1700 or 1800 years ago, there was an empire and religious missionaries that spread the script across Western and Central Europe, and then 500 years ago, new empires and more missionaries spread the script further. So today, there are over 400 living languages on six continents which use the Latin script.

In the first expansion, a few additional letters were added to Latin's original 21, but mostly, new phonemes were dealt with using diacritic marks. And in the second expansion, diacritic marks were added to an even greater degree, so there are now some 20 different diacritic marks that are used in one language or another. Next slide, please.

We have these seven panel members who have been involved over the course of the project. When we started out, between all of us, we had at least a passing acquaintance with most of the European languages that used the Latin script, which gave us a start at dealing with the whole thing. Next slide, please.

As you can see, we developed some criteria for which languages we would actually include in the repertoire we developed. We included any language which is the official language for either a nation or for a province, state or region within a country, plus, we added languages which are EGIDS 5, which is to say they're not an official language, and we took those when we had more than a million speakers. And that got us down to only a little over 200 languages.

We did this in the first couple years of the project, but it should perhaps be noted that just this past September, we found a couple of more languages which also meet that criteria which we had missed initially,

so it will not be astonishing if the public comment period turns up a couple more.

While we now have some 200-plus codepoints for the Latin repertoire, any given language uses less than a quarter of those. Frequently quite a bit less than a quarter, which means that users will have never seen a lot of the possible codepoints for the Latin script.

As noted, we have some 20 diacritic marks that are used, but any given language will only use a few of them, and as a rule of thumb, if the user is acquainted with a half dozen different languages, he may have seen as many as half of the diacritic marks.

Just to make things more interesting, Latin also makes extensive use of different fonts, and so users have all been well trained to ignore small differences in the shape of letters, which means if someone is expecting a particular diacritic mark or a particular letter and they see something that's different that they haven't seen before, the tendency is to assume it's a font difference rather than realizing that that's a difference that actually is important. So there's lots of opportunity for confusion here.

The Latin GP has used a very narrow definition of what constitutes a variant, on the assumption that for new TLDs, the similarity review panel will catch any conflicts. In spite of that, we have ended up with a lot more variants than most alphabetic scripts have, and we have an event larger number of what we have referred to as confusables, which are pairs of codepoints which some of the panel members—even in some cases a majority of the panel members—couldn't distinguish even comparing them side by side, but they didn't quite meet our very strict

definition of what was a variant, so they are simply included as an appendix.

I would note that as Pitinan mentioned earlier, IDN recently published guidelines for second level domain names, and the guidelines talk about variants and how to handle them, but for second-level domain names, the manual review, which is incorporated in the TLD process, isn't feasible. The IDN project may wish to give some serious consideration to whether those confusables that we've identified should be considered variants for the purposes of SLDs. Alternatively, of course, we can just decide that DNS abuse isn't something we care about and ignore the issue. Next slide, please.

We have gone through a number of versions of our document and gotten feedback from the integration panel each time. Most recently, in December of this past year. Next slide, please. We're now just about finished responding to the IP's various comments and putting together the test datasets, that sort of thing. We hope to have this ready to submit to the IP at the end of this month or early next month. Next slide, please.

To summarize, we hope to get that to the IP in April. Assuming there are not very many substantive issues raised, we optimistically hope that in early June, we can release the report for public comment. Assuming that there are not too many substantive issues raised in the public comment, we would hope to be ready to publish the final document in September. But that of course is a best case scenario, and I notice that the Korean generation panel went for public comment once and then

came back and worked some more before finally getting to their new version and going out for public comments again. And so it's entirely possible that we will see the same phenomenon. I think all the members of the panel hoped and pray that it doesn't come to that, but it's certainly a possibility.

If there are any questions, please let me know. Thank you.

PITINAN KOOARMORNPATANA: Thank you, Bill. Sarmad, do we have questions in the chat?

SARMAD HUSSAIN: Yes. Thank you, Pitinan. Thank you, Bill. We have one question in the chat. The question is by Sivasubramanian. The question is, in what scenarios does a registrant apply for a cross-script domain name? Does the registrant choose four alphabets from an Armenian keyboard, buys a Greek keyboard, choose three alphabets from Greek, and puts together a domain name to be registered? Please explain why there is so much attention on the extra script variants.

BILL JOURIS: I don't know how to answer the first part of that. The reason for the attention to the cross-script variants, or variants in general, is as Pitinan explained earlier, when the users can't tell the difference between the symbols that are used in a domain name, it causes security issues, it causes usability issues, and that's why we're paying so much attention to it.

The process part of your question, I simply don't know how to answer. Perhaps Sarmad can help us out.

SARMAD HUSSAIN: Thank you, Bill. I think you have been able to answer the question. I don't see any more questions in the chat. Thank you.

PITINAN KOOARMORNPATANA: Thank you, Sarmad. All right. Thank you, Bill. Then let's move on to the next generation panels. Yin May Oo, please, the floor is yours. Please also speak a little bit louder and also slowly. Thank you.

YIN MAY OO: Hello, everyone. This is Yin May Oo from Myanmar Generation Panel, and I am the co-chair of the generation panel. Our chairperson is currently in Myanmar and during the Internet blocking time, she cannot attend the meeting. So I will continue with the presentation. Next slide, please.

This is the agenda for today. Next slide, please. So from Myanmar GP, we cover the Myanmar script. Myanmar script is shared by many languages, mainly the language users with IDN Myanmar, but there are also many language users across the world. And majority use the Burmese language as it is the official language of the country for many years, and we have Shan language, Rakhine, Sgaw Karen, Mon and Pa'O Karen. So this is the majority of the ethnic groups that use the script, and they're all above EGIDS 5. Next slide, please.

So, as you can see, we have many languages sharing the same script. Our country is geographically located at the east side of Bay of Bengal, and along the river, we have started to use the script since [1811-12 year,] and we have developed since then. So our language, our script based on Devanagari, and we adapted the character sheets from [Hmong] tribes who are living in the Andaman Sea area in 11th century. So we have influence from both north and south tribes.

So along this Irrawaddy River, we have flourished the culture and language through the years. And around 600 years ago, we have improved the use of the language. So from the beginning, we just wrote down [first tongue and second tongue,] and then we have adapted to use more diacritics in [inaudible] era, around [1718]. So after that, we have adapted to colloquial Burmese.

So we have so many languages, but we share most of the characters, most of the consonants and diacritics. So in total, there are 99 codepoints that we have used in MSR. Next slide, please.

So for the nature of Myanmar script, in most of the Devanagari languages, we have got one consonant, and it may be surrounded by forward diacritics like [inaudible] left right around the center consonant, but what is more special for our language is that we can use another consonant as the [vowel enhancement] like [inaudible] so that we can enhance the sound, and then we can also add more diacritics as [inaudible]. So we can take like two or more [inaudible] to pronounce one pronunciation, and since we have a lot more complicated diacritics, we have a to take care of the rules.

And moreover, we have [syllable chaining,] which means more than one pronunciation can be combined to one word, which this kind of writing is before [1180], but some of the words, we keep on using into now, so we cannot leave it behind. So we have to consider all the writings from the perspectives of different languages and make everything adaptive in one [inaudible]. Next slide, please.

So we have analyzed and considered many things for in-script variants. We have single-character combinations, single-character which looks identical to the three-character combinations like you can see in the table. So we have analyzed through the whole script and from all the languages. We have to consider what could be the variants in the script.

So for the variants, row number four, five, six, seven, they are not only homoglyph, they could also be homophones and they also have the same semantic meaning. So we added in the rules that these characters should not have been in the same label, so that [inaudible] or cross-script analysis. Next slide, please.

We have more than three languages that share the same character, so from row one to six, these are the [inaudible] that we have analyzed through, and we have similar characters in Georgian script and then we have Oriya letter, and we have Oriya diacritic. We consider this because most of the Myanmar characters are based on the circle shape, the O shape calligraphy, so comparing with many other scripts, we just list out that these are the possible combinations which can clash with Myanmar script.

So this is our cross-script, so this is the result that we come out after checking all the different languages which could have similar [inaudible] to Myanmar, and the rest are just listed in the confusables. Next slide, please.

So this is just a bigger view of what we consider confusable, because the locals may [rate] these characters the same since the difference is very little. Next slide, please.

So those are the updates that we have currently. Our GP is formed in 2018 June, and our last update before this meeting is about 2020, so we hope to release for public comment in the middle of 2021, like by end of May, and finalize LGR before the end of 2021, hopefully in September. Thank you very much.

PITINAN KOOARMORNPATANA: Thank you, Yin May. Now we have a few minutes to the end of the sessions, and it is open for question and answer. Is there any question or comments to read out, Sarmad?

SARMAD HUSSAIN: We don't have any questions in the queue at this it me.

PITINAN KOOARMORNPATANA: Thank you. I saw some active discussion in the chat, so if any of you would like to take the mic [inaudible] possible. Okay. I don't see that we have further question or comment to discuss in the sessions, so if you have any further question or comment, please feel free to reach out to

us. The e-mail address here, idnprogram@icann.org, and we can respond to your question.

So I guess with that, with all the speakers, do you have any final remark before we close the call? All right. Hearing none, then I guess we can close the call for today. Thank you for joining, for participating in the session and for your attention, and also for the discussion. So we can stop the recording now, and please have a good rest of your day. please take care. Bye.

[END OF TRANSCRIPTION]