MARRAKECH – IDN Program Update
Wednesday, March 09, 2016 – 10:45 to 12:00 WET
ICANN55 | Marrakech, Morocco

UNIDENTIFIED MALE: This is the ICANN55 IDN Program Update meeting on March 9th, 2016, at 10:45 in the Ametyste room.

SARMAD HUSSAIN: Okay. Thank you very much for joining the session on IDN Program Update. Let's get started.

Next slide, please. Thank you. Here's an overview of our session presentations today. We'll give a short overview of different projects and programs being conducted under the IDN program currently. I will go in detail in some of those in the first presentation. However, some of the details will be presented separately as well. We will have an Integration Panel update by Marc Blanchet; IDN Implementation Guidelines work going on by Edmon Chung, who's the Co-Chair of that working group; and an update on work currently going on on Reference Second Level LGR development by Michel Suignard.

We also have updates from community members on the different LGR Generation Panels, and we will have updates from

the Khmer Generation Panel by Rapid Sun, from the Lao Generation Panel by Chittaphone Chansylilath, and on the Latin Generation Panel by Chris Dillon.

Then we will have a question and answer session after that.

Next slide, please. Onwards to overview of the IDN Program. Next slide, please. Okay.

The IDN Program is currently undertaking multiple projects, and these projects we normally divide into two categories: one set which is looking at top-level domains, and then another set of projects which is looking at second level for gTLDs. Then we have a separate outreach and community involvement thread, which we undertake to apprise everybody on the progress as well as engage them in the work.

I will be presenting updates on some of these projects, except those for which we have community members and other members here who will provide more details on specific projects they are working on.

Let's move ahead. Next slide, please. First of all, I think one of the larger projects we are involved in is developing label generation rules for the root zone – excuse me – and this project has been active since 2001, where the community has been trying to determine what is the best way to define valid top-level

domains in different scripts and determine their variance. It has multiple parts to it, one part which is developing linguistic rules for developing these labels and variants for each of the scripts. In addition, we have a project which looks at implementation details of these variants.

Finally, we are also looking at a tool to use this data. We'll be talking more about that as well.

We are happy to report that the first version of label generation rules for the root zone have been released on the 2nd of March of this year. Two communities finished their work and handed their proposals to ICANN for integration in July of 2015. We took the proposals for public comment, as per the process, and based on the public comment, the communities finalized their proposals for integration in November of 2015. The Integration Panel undertook the evaluation of these proposals and publishes the first version of LGR for public comment, again, as per the process, in December of 2015. The public comment closed on February 2016, which led to the eventual publication of the first version of LGR.

The two communities which finished their proposals included the Armenian and Arabic scripts. The Arabic script proposal was integrated, but the Armenian script proposal, though successfully evaluated, was not integrated because it had some

cross-script variance with the Latin, Greek, and Cyrillic Generation Panels.

The Integration Panel thought it would be wise to wait until those Generation Panels have finished their work and there is clarity on how they look at the same problem so that a common solution can be determined. So the Armenian Generation Panel LGR proposal was not integrated at this time.

Next slide, please. Here is the status of different Generation Panels. We have many communities which are now active. As you can see, the Arabic and Armenian have already finished their work. We're happy to report that the Khmer and Lao Generation Panels have also completed their proposals, and they're currently informally interacting with the Integration Panel to get any feedback on any gaps which are still left before they go for public comment. They will be going to public comment soon, and we'll be hearing from the Lao and Khmer Generation Panels in more detail today as well in this session.

In addition, the Chinese, Japanese, and Korean Generation Panels are all seated, and they continue to work within their communities and also coordinate efforts across them to find common solutions since they share Han script across the three Generation Panels.

The Ethiopic and Cyrillic and Latin and Thai Generation Panels are also seated and making good progress. We are also working with other communities to get their Generation Panels started. So this just gives a brief overview of where we are in the whole process.

Next slide, please. We have now all this data which is being produced through all these Generation Panels. One of the projects we are undertaking is basically to have a very formal, machine-readable mechanism to store this data and process this data.

There has been a project which started to develop formal specifications to store this data. The specification is being developed by the IETF through the Lager Working Group there. The working group actually had their last call last week, and that was finalized. We'll have more updates on it by Marc Blanchet as well. That work is now also coming to a closure, so we'll have a Standards Track RFC representing this formalism.

Based on that formalism, we're also developing a tool which the c0mmunity can use. We have actually already developed three phases of this tool. This tool is now available through the ICANN website. It's called LGRTool.ICANN.org. It does require credentials to sign in, and we are more than happy to provide those credentials to you. If you want to access the tool, just e-

mail us at IDNProgram@ICANN.org, and we'll provide those credentials.

This tool allows anybody to use the online tool to create an LGR. Also, if an LGR already exists, you can upload that LGR and then use that LGR to determine if a particular domain name is valid, based on that LGR, and what are the variants of that label, given that LGR. It also allows you to take two different LGRs to determine what the differences are between those LGRs. Or you can create a union or an intersection of those LGRs. So it allows you to manage LGRs as well.

We are now finalizing another phase of this project in which you will be able to also upload an existing set of labels to see, if you change a version of LGR, how that impacts the existing labels in the system.

Finally, once this application is completed and tested, we intend to release this as open source for the community to use this application and integrate it into their own systems, beyond, obviously, its availability through ICANN.

Next slide, please. Here are some screenshots of these tools. For example, here is a screenshot of the first phase, which allows you to enter an LGR code point. The reason for this work was that the XML specification is rather complex, and it may be hard for Generation Panel communities and other community

members to develop the XML itself. So this gives a very easy user interface, where you can just do one code point at a time, write rules, and define variants in a very easy-to-use manner. That just automatically creates the corresponding XML at the back for you to use.

Next slide, please. Go back one. Oh, it's okay. Yeah. This is the second phase, which, once the LGR is loaded, you can see there was a label which was typed in the text box there. It tells you that the label is valid using the LGR. If you look down a bit, it says there are 1,600 variants generated using this label, of which 1,542 are blocked, 50 are invalid, 7 are allocatable, and 1 is valid. You can then download the corresponding file, which lists all those variants and their dispositions in more detail. So it provides you all the details of variants.

Next slide, please. Thank you. This is the final slide, which shows the third part of the functionality, where you can see that you can load two different LGRs and you can do a union, intersection, or difference of those LGRs if you're managing LGRs at your end.

Next slide, please. We are also undertaking the fast track process for IDN ccTLDs. Currently there are 49 strings which have been successfully evaluated, and all of which represent 39 different countries and territories.

Next slide, please. Of these, 43 have been delegated, again, which represent 33 countries and territories, and that overall represents 18 different scripts and 27 different languages.

The fast track process is currently under annual review. Basically, there's a ccNSO Working Group, which is working on looking at public comments which was received when this annual review was opened. Based on their feedback, the process will eventually close.

Next slide, please. As far as the communication efforts are concerned, we have been very active in updating the IDN web pages at the ICANN website, and you can access it very easily by just going to ICANN.org/IDN. We have been presenting updates all the ICANN meetings to the community, and we also go to individual SOs and ACs to present updates to them.

We've also been active. We're reaching out to different communities to engage them in the program. We held a workshop on IDNs and African languages in Congo last year to engage the community from Africa to get involved in the Ethiopic and Latin Generation Panels. We also held a training for the Korean Generation Panel in Seoul, and we assisted the process of public workshops for Khmer and Lao root zone LGR proposals for their respective GPs. They'll talk more about that later today.

We also maintain an e-mail list and we [keep pages] to share the promise of our work with the community.

Next slide, please. We are reasonably easy to reach, so come visit us at ICANN.org/IDN. If you have any questions or queries or need any more information, just simply write to IDNProgram@ICANN.org.

So that's the update from the IDN Program for now. Let's move on to the next sessions. The next presentation is by Marc Blanchet on an update by the Integration Panel. So, over to Marc.

MARC BLANCHET: Thank you, Sarmad. Next slide. This is a short report of Integration Panel activities since the last ICANN. We reviewed the final Arabic script LGR and the Armenian script LGR. We produced the first root zone LGR, named LGR-1. Public comments are done. We did interactions with the current active GPs, such as the Armenian, CGK, Khmer, and Lao. We reviewed the new GP proposals – Ethiopic, Cyrillic, and Korean – and we formalized an integration processor for root zone LGR, the last point here being that we prepared the work, as all the GP LGRs will be integrated together, and that was requiring some processing, I guess, of the files.

Next slide, please. For the LGR-1, we received the Armenian and Arabic LGR, who both passed public comments. We reviewed the Armenian LGR and found that it needed to be considered together with other scripts because of the cross-script variants. So it is deferred. Therefore, as soon as the other scripts are in, then we will be able to integrate all of them together.

IP also reviewed the Arabic LGR and found no dependencies with other scripts, so it has been accepted inside the first LGR of the root zone.

Yeah. Again, we conducted an extensive review of the integration process. What we did is actually took current prototype LGRs of various other scripts and then just tried them all together so that we are ready when the final versions are. So it's more a technical work on our side for the integration itself.

Next slide. Those are, for your reference, various links of the actual files of the LGRs.

Next slide. On the IETF side of this, as Sarmad was pointing out, the Lager Working Group was formed to finalize the XML format to be a Standards Track RFC, so fully reviewed by IETF.

Actually, the slide now is a bit obsolete because it's not converging to final specifications. It has converged to final specification, and, instead of expecting to be sent to [IAG] for

approval next month, I actually pushed the button yesterday to request [IAG] for a formal review of the specifications. So we were prudent when we made the slides, but we were making more progress. So now it's in the hands of the [IAG], so should it be in IETF last call in a few weeks. Obviously the IETF meeting is coming soon, so it may be before or after. If everything goes well, we may have an approved specification before June.

IDN 2008 tables are not updated, and this is not news, just considerations that influence the work in the future related to an issue that was found and the consequence is the current IDNA tables being based on Unicode 6.3, where Unicode is almost nine now, and ten in a year. That means that any new code points after 6.3 cannot be reviewed or used at the moment.

Next slide. What we foresee in the future? A review of Khmer and Lao LGRs, and maybe Thai, too, is coming, and an interaction with the CGK Generation Panel, concerning their LGR address. They have a difficult problem because they need to coordinate together because of the various variants. So it's a lot of work, and they've been doing a lot of work. We actually interact with them a lot to help. And obviously, a review of new Generation Panel proposals as they come. Thank you.

SARMAD HUSSAIN: Thank you, Marc. We'll move onto an update on the work which is now going on the revision of IDN implementation guidelines. I will hand it to Edmon Chung, who's the Co-Chair of that working group.

EDMON CHUNG: Thank you, Sarmad. Edmon here. I guess we'll move to the next slide. Basically what we'll cover, I'm going to give a little bit of background of what the IDN implementation guidelines are there for and what the current iteration is working on. I'll talk about the composition of the working group, and then some of the current topics that the group is considering.

Next slide, please. This is a very special piece of document, I guess, in the ICANN set of documents in terms of policies and implementation. This is called the ICANN implementation guidelines, but as we finalize it, it will have immediate implications in all the new gTLDs, at least that offer IDNs, because this is a document that is included in the contracts with the new gTLDs. This is a piece of document that is designed to address a number of IDN implementation issues to make sure that registries are both compliant with the guidelines but also to take into consideration user confusions and other types of concerns.

I see here that, for ccTLDs, I actually have a different understanding of it. It says, "Recommended for IDN ccTLDs." In fact, I think it's required by IDN ccTLDs. I was in the working group that created the IDN ccTLD fast track, and it is pretty clear that IDN ccTLDs must comply with the IDN implementation guidelines.

This is the only document that covers both gTLDs and ccTLDs that I know of at ICANN. This is a background of the IDN implementation guidelines. In previous work at the GNSO, especially as we work through the New gTLD Program, and as the IDN implementations evolve, there is some need to revise the guidelines.

In fact, these guidelines were set out, I think, first in 2006 or 2007, and it's been through three or four iterations already as we see the implementation of IDNS, I guess, in the real world. Currently, we're going through a process to update, really, the guidelines.

Next slide, please. There was a call that went out to the community to participate in the IDN Guidelines Working Group, and there is participation from GNSO, ccNSO, ALAC, and SSAC. I'm trying to count to make sure. If I remember correctly, there should be six GNSO members, two from CCNSO, two from ALAC, and one active member from the SSAC but who's also on the

Board of the IDN Working Group, and, as I understand, Ram Mohan sitting next to me. And then Patrick from SSAC is also, I believe, on the mailing list, and have committed to the work coming out of the working group.

Currently, myself, from the GNSO, Mats, from ccNSO, and Wael, from ALAC, are the Co-Chairs for the working group.

Next slide, please. The next couple of slides really are the six topics that we have identified that should be incorporated into the next version of the IDN implementation guidelines. I don't think I will go very much into detail about it, but in the overall process, we're hoping that we will put out these six items. This is the opportunity to let people know this is what we're doing and try to get a little bit of feedback. But for the formal feedback, we're looking at an initial report, hopefully around the next ICANN meeting. That's still the current target.

Then, based on the formal process for the feedback from the public comments, we will work on a final report, which would, in effect, eventually update the IDN implementation guidelines.

The six items. One is, those of you who were in the UASG Universal Acceptance session earlier, Andrew pointed out a point about the IDN standards. The protocol standards have changed between the 2003 and 2008/2009 timeframe, and

there's still some lingering transition issues that continues to be important. So we think that's an important aspect.

Over the many years that IDN has been in discussion, many terminologies were proposed and created and morphed over time. One of the hopes is to at least gather some of the key terminologies and try to bring some kind of, I shouldn't say standard, but at least some sense for people to follow the discussion.

The format of the IDN tables is another place we're discussing as well, but we're going to incorporate some of those into the new IDN guidelines. Consistency of IDN tables across different registries, this is going to be one of the issues that we will look into.

Next slide, please. IDN variants. Of course, from the very beginning in the first iteration of the IDN implementation guidelines, IDN variants were touched on, but especially over the last five years, a lot of work has been done from the community and from ICANN on IDN variants. So there's a lot of updates that may come into the implementation guidelines.

We'll also need to touch on the confusability and similarity issue. Again, this is from the very first IDN guidelines. This has been there, but it has been refined over the years, especially in light of the new gTLDs as well. So that's another item.

Finally, registration data, what is commonly called the WHOIS or contact data for registration, that seems to be an area that should be included as well.

So those are the six broad strokes that we're looking at. Again, this is not a formal initial report yet, but any input at this point or at the formal processes is much welcome.

Next slide, please. So that's where we are. We do have a face-to-face meeting here, and those are the links about the program. You can follow our work from the community wiki as well.

Thank you.

SARMAD HUSSAIN: Thank you, Edmon. We go onto a short update on the reference second-level LGRs. It will be provided by Michel.

MICHEL SUIGNARD: Okay. Next slide, please. This will be the outline of the presentation. We'll go through the background of what it is, and then go to different subject of second level LGRs, which are guidelines [inaudible] to public comments on the status of the LGR.

Next slide. The [inaudible] to explain what it is, in fact the work is to develop a set of LGRs covering selected languages. This is

unlike the root. We're talking languages here, so it's a probably bit more aligned with what was done on the second level, where you have seen these IDN tables that were addressing languages.

In fact, at the same time, we're also developing a guideline to explain all those LGRs being developed, the process to collect information.

The scope is in fact 29 languages. That's different than what I had. There's 29. Latin has 16, then Cyrillic has eight, and then there's two with mixed scripts. By that I mean that the Japanese and Korean in fact do include Latin. That's a big distinction between the root and the second level. We see in the second level you may allow ASCII in the DH, including the hyphen, which are not allowed in the root level.

For Japanese, unlike the root, we allow [inaudible] or Latin characters, as well as for the Korean. For the Arabic, we also have Arabic, which is single script. We see on Chinese also Latin. Hebrew is a single script.

So that's the scope. Again, this is when we're talking about the list, where Arabic as a language, not as a script, which is different, again, from the LGR-1 work.

Next slide. The guideline was really how we get to those set of LGRs. The key aspect of how you define code points that you

need for languages is not that simple, in fact. Some of them you have really to go at different references. Some have well-established references. You could think maybe of the French. The [people] have the French academies that would give some idea on what is supposed to be used to write French.

For other ones, you have to look at, what was there previous practice in my IDN tables? You may also have to look at the different publications. For example, we found that, in other countries, there were some publications that were done to determine what was [coverage] for various European languages.

For Germany, in fact, you can find some good references. For others, you really have to look around. We in fact looked, for example, at CLDR, which is work that is done by the Unicode Consortium that collected information related to languages or writing systems. It is in fact pretty comprehensive. It's been used by a lot of various platforms, very successful platforms, in the IT industry.

So we see the guidelines define how you find those sources, but also define how the process is done. So there's a multiple-step process where you do first create a set, and then you have to go for reviews, both linguistic review by linguistic experts. You also look from a security and stability point of view, which we see as very important in the end. The last stage is basically a public

review of the repertoire. So these guidelines basically describe the whole process and has been presented by ICANN for receive feedback on those.

You published a new version now, right, of the guidelines? And now the guidelines is available on the ICANN website.

Next slide. When the guidelines were put for public comments, we got some comments that were not mostly about the process but more about the project itself. There were basically some concerns on how they were going to be used. That was a bit outside of my scope, frankly, as myself developing the guidelines themselves because we were describing how to create the LGRs, not how they're going to be used.

So I think that some explanation or clarification were provided in the answer to the public comments to basically explain again how those LGRs are going to be used. My understanding – obviously, again, that's a bit out of scope for me – the scope is we need to be using pre-delegation testing – that was my understanding – [on] only a reference point. They basically established a reference for anyone that is trying to do IDN language-based tables, so they know there's at least a base point so they can expand on it if they think it's not enough.

And they can always evolve. These are not necessarily totally 100% stable. They can be augmented if people think that there

needs to be some additional code points or additional rules to be added to those LGRs.

Next slide, please. Next. Okay. Challenges, that's more on the technical work of doing those LGRs. Like I said before, language coverage is a moving target. It's not that easy to determine what is used on a given writing system. There's a multiplicity of sources, and sometimes there's so many of them that you don't know which one is the right one. And they do diverge sometimes, so you have to make a judgement call on what is the right answer.

Then there's a dictionary. They move depending on what's the level of acceptance of foreign words. This is especially true even in the Latin and Cyrillic systems – let's say you live in the Ukraine – on how much of Russian you want to except in your Ukrainian writing system. Same in Yugoslavia. From Yugoslavia, you have a lot of difference between the different section of what used to be known between Serbia, Montenegro, Macedonia, and all those places. They have slight variations but sometimes they have influence from the neighboring entity.

So we have to take that into account, and see, for example, in each of the systems, especially between those of Cyrillic and Russian, if you want to include Russian code points into a non-

Russian language. This kind of decision you have to make. To some degree, you have to make a judgment call.

To a large degree, we've been pretty conservative. The principle of the work was to be conservative. So we could obviously add later, but at least as a first step, we considered it to be more on the cautious side then really on the liberal side.

Then there is there is scope of language. For some, it's pretty easy. If you want to define Greek – Greek is not part of this one – but let's say you want to define even Finnish. You have a pretty good idea of what that is. But if you try to determine Arabic as a language, we took as an assumption that we would go, for example, for a pretty expansive view of Arabic on that aspect. We went from basically Morocco to Iraq, from the west to the east, and then from the north. We went from Syria on the south to Sudan, I guess. That's probably where you would consider Arabic to be used. So we went pretty far.

At the same time, it's a bit complicated to determine. For example, in Iraq, you do use some Persian letters because there's some brands in fact that do need characters beyond pure Arabic to be represented in Arabic countries because they do use different consonants. So you have to at least consider those.

So you have some judgment calls to do all those things, especially for some that are very widely used. Even the English

aspect is also problematic because English has a lot of borrowing from, for example, French. A typical word is “naïve.” Does “naïve” have an umlaut in English or not? So, anyway.

Opportunities that we see, we’re also using the fact that we have done a working LGR. For example, there’s an Arabic LGR-1. That useful for us when you do a second level. In fact, there was a lot of work that we could use, especially on the variants side. All the variant work that was done by the Generation Panel for Arabic was very useful for doing work on the second level.

On this, obviously, work was done on second level at DNS. A lot of tables exist already. They’re out there, and that’s very useful as input. Plus, the work that was done by the IIS Group, which was the first one, in fact, do this kind of second level LGRs.

Next. Status, I said that they were all ready for public review, but in fact – so the work is done, but it’s now inside ICANN, and ICANN is doing internal processing to validate them. So maybe later Sarmad will have more to say on those.

On the XML files, we see we’re talking LGRs, so it’s a set of XML files. But they have a very deep documentation built into the XML, so they do describe themselves, what they’re about, on variants, on the repertoire, and on why characters were added or not added.

We have a documentation, in fact, that does extract automatically for the next XML. That way, we ensure that we're always synchronized between the documentation on the XML itself.

Documentation elements as a source: references, repertoire, rationales, and variants. [inaudible] some of them [are] rules, especially in Hebrew and Arabic. They need a set of rules where, in Latin or Cyrillic, you don't really need rules, except for [inaudible].

Next, please. So you can see already the guidelines. It's on [9]. Obviously we're using the XML Lager specification as well. It's probably one of the implementations of LGRs at this point because we have 29 LGRs that were done on this project. So we have been using the XML spec to its full extent.

Obviously, I'll be open for questions later. Thank you.

SARMAD HUSSAIN: Next slide, please. Thank you. We'll now move onto community updates.

RAM MOHAN: Just a question for you. On the last slide, you were talking about language convergence, etc. If you look at the African region,

there are a whole lot of languages that are now standardizing and using the Arabic script. In Unicode, it's going to treat it as an extension, if you will, to the Arabic code page.

What happens if those languages want to be represented in LGRs or in the DNS? Because Arabic is itself covered now. But you have I think it's called Ajami, and there are other representations. Many of the languages here in Africa have now chosen to use Arabic script, but they're not using just a standard Arabic script. So what would be the path for those [representators] from those languages?

MICHEL SUIGNARD: Okay. You could either do as separate languages, because after all, they're all separate writing systems. So you could follow the same principle that is done on the second level project. Now, you will have one for each of the writing systems, only there's many, I understand, that would be maybe restrictive.

The other way is you could in fact do some sort of Pan-African Ajami if you want. You could do such a project if you want to. But at the same time, you could use in fact the Arabic LGR-1 itself and make it Pan-Arabic. There's nothing that prevents languages to be merged in a single project. Do you see what I mean?

You could, for example, do a Latin set that will cover all Europe, the same way you could merge languages together and make a single table for a set of languages together. There's no requirement to separate languages per table. You could have a table that covers multiple.

So if someone wants to propose an Ajami Arabic from Western Africa – it's beyond me to some degree – someone can do that. Either ICANN or whoever can basically support or sponsor such a work. Yeah.

SARMAD HUSSAIN: At top-level, obviously there is work which is already done. For second level, we have members of a Task Force on Arabic IDN Working Group, which is a community-based effort. They are actually one of the second things after the LGR-1. They have actually undertaken work on developing Arabic script tables for second level. So this is work which is now being taken up by the Task Force on Arabic IDNs, and they continue to do their work as a community group to represent Arabic script through IDNs. So that may actually be another effort which is relevant in this context, if you are looking at it from a script context. If you are looking from a language-based context, then obviously one would need to develop each of these tables separately for languages.

In the interest of time, I hope we can continue, and then we can get questions.

[ALAN GREENBERG]: Apologies. I do have to go, but I think, though, there was another question that Ram was asking, which is about new characters that are introduced into Unicode. This was I think something Marc was talking about a little bit earlier, that there is this problem that we're not sure what to do about the rules. I don't think we're going to solve that here, but there is an urgent problem there. I think contributions to that issue would be welcome.

SARMAD HUSSAIN: Okay. We'll come back this discussion at the end of the presentations. Let us move to the update from the Khmer Generation Panel by Rapid Sun.

RAPID SUN: Good morning. My name is Rapid Sun. I'm working ISO member [inaudible] of the Khmer Generation Panel. Please, next slide. Next slide.

I will update you about our progress on the Khmer proposal of the LGR.

Next slide. The Khmer language has been around since the 7th Century. We also somewhat borrow from Sanskrit and Pali. We are using for our citizens, about 15 million people, and some parts are also using in Southern Thailand, and also in Vietnam.

Please, next slide. Next slide, please. There are about 146 characters in our alphabet [inaudible]. Also we write from the left to right. Some languages, like Thai and Lao, derive from the Khmer language also.

Next slide. At the Khmer Generation Panel, there are about ten people, and the members come from the technical side and also come from a linguistic side and also come from academia.

Next slide, please. This is our proposal that we submitted to the Integration Panel. We plan to develop the code points and the variants, the labels, and the rules. We finally finished the public consultation workshop on the 15th of February, just last month.

Next slide, please. In our code point development, because some of our members come from the technical side, like the networking skills or from the ccTLDs, many of them are not skilled in Khmer language. Then we're doing with a smaller group with only linguistic people [inaudible] linguistic. Then we defined the code points, like consonants, dependent vowels, independent vowels, and some [inaudible].

We also developed the variants because there is one variant from Khmer script when we're using the subscript. Also, we defined the cross-script variant between Khmer and Thai, between Khmer and Lao and Myanmar.

Next slide. Totally, because this is [inaudible], we have developed about 12 label rules. The letters is 12 label rule and with also the XML. For the cross-script variant, we had a meeting yesterday with the Lao and Thai and with also the English and [inaudible] The character is similar, but when we form a label, it is not similar, so we don't consider it as a variant.

Yeah. Next slide, please. Until we develop the proposal, there are about ten additions to the proposal. Then we share the proposal, when we develop it, with the Khmer linguists. We share with the other members in the community for the feedback. Also, we share the proposal with the Integration Panel. We have maybe four or five feedbacks already. We host a public consultation, which we invite people from universities, from private sectors, from NGOs, and from Civil Society who are interested in working on the ICT in Cambodia.

Yeah. Next slide. This is our update of the Khmer Generation Panel. Thank you.

SARMAD HUSSAIN: Thank you, Rapid. We'll move onto the update from the Lao Generation Panel, and we have Chittaphone here to do that. Over to you, Chitta.

CHITTAPHONE CHANSYLILATH: Okay. Good morning, everyone. My name is Chittaphone Chansylilath. I'm from the Ministry of Posts and Telecommunication of Lao PDR. I'm a member of the Lao GP. Today I'm going to present about our work and also our GP, the components of GP, and also some work that we have done.

Next slide, please. Okay. I will start with the Lao Generation Panel. The Lao Generation Panel has been approved and stated by ICANN on the 15th of September, 2015. There's 13 members in the GP, which consists of three linguistic experts and four from IDN operations, one policymaker, and one from media, and the last four from the Lao Localization Project, which is to develop Lao language in computer systems.

The GP is chaired by Mr. Phonpasit Phissamay, who is the Director of the E-Government Center. He's had expert with Lao language development since 2003, and also he's the one who integrated Lao into the E-Government system.

Next slide, please. Next. Sorry. Okay. Next I will move onto some introduction about Lao. The official language of Lao originated

from the Tai-Kadi language, which has maybe the same root with Thai and Khmer. This language is spoken by approximately 30 million people, which is mainly in Laos and the northern part of Thailand and some areas in the neighboring countries, like Cambodia, China, Myanmar, and Vietnam.

Lao language is a tonal language, which, according to the book, we have six tones: the high normal, low normal, mid high, falling, mid falling, and low rising. The dialect is differentiated into five main groups: Vientiane, Luang Prabang, Xieng Khuang, [Khammuan and Champassak]. The script is derived from Pali and Sanskrit and has been continuously developed time to time until because it's become unique to Lao language. There is no space between words and we write from left to right.

Next slide, please. Next I will conclude a bit about some work that the GP has done. So far we have determined the principles to be used to determine the valid code points. We have analyzed the variance between in-script and cross-script variants between Lao, Thai, and Khmer. Also, we have identified the WLE rules for the root zone, and also we have written the proposal, which is now under review by the IP. We plan to submit the proposal really soon.

Next. This is the code point that we have identified. As you may know, we cannot put our [inaudible] Unicode table into the root

zone. So according to the regulations from ICANN and also IDNA 2008, we have in total 51 correct to be included in the root zone. [inaudible] 27 consonants, 18 vowels, 4 tone marks, and one sign.

Next slide, please. This is what we have researched for in-script and cross-script variants. In some code points in Lao, we can write in a different sequence, but it looks totally the same, as you can see from my slide.

So in this, sequencing is not consistently supported by all the [inaudible] system. Therefore, the Generation Panel has decided that only the valid sequence should be considered, and the less is we just make spelling [inaudible]. This is the conclusion from public comment and also from the GP. So in conclusion, we have no in-script variant in the Lao language.

Next slide. Next slide, please. Okay. For cross-script variants, we have done research about between Lao, Thai, and Khmer. This is my own slide, so normally for Lao and Thai we have 12 pairs, which is more similar and only one consonant which is similar. But because we just did the public seminar last month, on the 18th, after the comments and after consideration, we said that the Lao and Thai can recognize which one is Lao and which one is Thai. After the consultation yesterday with the Thai

Generation Panel, we concluded that we will not consider as a cross-script variant between Lao and Thai.

Next slide. For Khmer – okay. Can you move onto the previous one? For Khmer, we have only two characters which are more similar, but those two pairs are not consonants, but it is vowels. The vowel cannot be formed a label. So also in conclusion we have no cross-script variants between Lao and Khmer.

Next slide – oh. Sorry. Okay. For this one I want to explain a bit about how we can determine the WLE rule. This is the structure of the Lao word. We have in total ten characters [in] one syllable, as you can see, from X0 until X10. X0 represents the vowel before, which can be in front of the consonant.

Finally, in total we have ten rules, which is categorized into four groups: for consonants, for vowels, for tone mark, and also for the sign.

Next slide. That's all for my talk today. Thank you.

SARMAD HUSSAIN: Thank you. We'll move onto our last presentation for this session. So over to Chris Dillon for an update from the Latin Generation Panel.

CHRIS DILLON: Thank you very much. Next slide, please. Okay. This is just a brief overview of the scope of the Latin Panel, the members we currently have, some resources used, and then something about the Latin part of the Maximal Starting Repertoire, and then basically, what next?

Next slide, please. We have drawn up lists of the languages using the Latin script – very long lists – and we've split those into large areas of the world. This is just an example of that sort of thing, largely the work of my colleague, Mirjana Tasic, in fact.

Next slide, please. Currently we have all together over 20 members, but I think probably only a small number of them, much less than that, are active in the work.

Next slide, please. This is a brief indication of the current status. We haven't formed a panel officially yet. If you look at this, it's a traffic light system. X means this is the thing that's stopping us. In fact, I wasn't aware that so much work has been done in the area of training XML. There's also a tool which may go some way to turning that particular stop sign into an amber light.

But probably what should be a stop light here is actually the linguists because the situation there is that we are strong in Europe, so we cover the Romance, Germanic, Slavonic, and some other languages in Europe very well. But once you set foot outside that area, there are certain brighter points. We have

some African expertise. We have Vietnamese and Swahili expertise.

But the summary is that we are really looking for either experts on individual languages or actually experts on – well, they were described as language areas. So just Central Asia, for example – really large. Or the Americas. These are all areas where we have little expertise at the moment.

Next slide, please. Perhaps we're getting a little ahead of ourselves, but we have spent some time with more major resources. I'm actually looking at Unicode, CLDR, ScriptSource and Ethnologue. We took the code ranges, which are okayed for Latin use within the MSR. This is sort of a typical picture, where you can see the Latin ones. We basically think, yeah, there are large languages using them.

We can use something called the EGIDS Scale, which is defined by the Ethnologue website. It gives you a good idea of whether that language is in current use. I think it goes from 1 to 9. The rule of thumb is that if it's between EGIDS 1 and 4, then that language will probably be covered by this work. If it's 5 down to 9, then that's less likely.

But we have already discovered one language where we'd like to break that rule, and the language is Esperanto because it's spoken by a lot of people – about two million people – but it's

EGIDS 5, and it has some code points which it needs to write itself in the normal way of writing. So, yeah, that's the summary of where that one is.

Next slide, please. What now? As you can see, we need to find more members, and that means via conferences and reaching out to people. Once there are enough of us, then we can form the panel. Then we're really continuing our analyses of the code points and also in related in scripts. Cyrillic is the biggest one, but also Greek, Armenian, and possibly other scripts. Then, beyond that, we're talking about creating the repertoire and the WLEs, writing the report, and submitting for review.

Next slide, please. The last point I would like to make is that we've made very good use of the community wikis, so if you're interested in our work and you would like to join us, then you can go in and you can find, apart from the normal MP3s, the longer versions of what I was just talking about: lists of languages, lists of code points, the draft panel formation proposal, and our notes on all the meetings. You can find all of that in there. So for anybody listening to this who would be interested in joining us, it's actually quite easy. You don't really need to listen through hours and hours and hours of MP3s. There is a shortcut.

Thank you very much.

SARMAD HUSSAIN: Thank you very much for the presentation. We have now some time for questions, so let's open the floor for any comments and questions from here in the room. Yeah?

CAITLIN TUBERGEN: Hello. This Caitlin Tubergen speaking on behalf of remote participant Matt [Stofberg]. The comment is, "The work raised a question on the concept of language table. You stated that you had a conservative approach to the language tables, which is probably true for some tables, but definitely not all. I have been able to review the tables, especially the Swedish table. It contains code points far beyond what is used for writing Swedish. If such a table is accepted, then you could as well go to script tables instead. The concept of language table is meaningless."

MICHEL SUIGNARD: I'll try to answer that. There is obviously a controversy on that. I'm just doing what ICANN wants me to do. So if, to some degree, if ICANN wanted to have a Swedish language table – and some people are asking for them – so I understand. The concept of language table is controversial, but I don't really have an answer to that. Some people would think it's pretty stable. Some think

they are not. It depends on how you take foreign words. To some degree, you use authorities, then you follow that authority to determine what to do.

Again, I don't have a good answer to that.

SARMAD HUSSAIN: Edmon?

EDMON CHUNG: On that note, actually it's very interesting. The language script issue keeps coming back. If you even look at Chinese, we use a lot of Latin characters. Is that part of the Chinese language? Chinese script? Neither seems to make sense.

But for this particular matter, I think we're looking at registrations of domain names, and that's really what's relevant.

Back to my comment, actually I have one comment and two questions, actually, both on the reference tables as well. My first comment is, I understand the background or the context of the creation of those LGRs, but I think not necessarily from the expert side but from the ICANN side, the economic incentives or economic implications to registries are important. The acquiescence, if you will, that reference tables may create – or something that I think ICANN needs to be aware of, not

necessarily “change everything that you do.” But that’s an important background in how those tables are created. So that’s a comment.

Two questions. One is, I see that the 29 tables, or at least 29 languages or scripts, have been created. And I wonder if there was been consideration of those tables in the gTLD context versus ccTLD context. Because if you compare the two, they can be largely different because, in a sense, especially going back to [Mat’s] point about language script and what’s relevant, the user’s perspective, or really the potential lack of any kind of indicator in the gTLD space, may not be the same as the ccTLD case.

For example, Arabic or Han characters. Arabic, as Michel mentioned, is used across different languages or areas. But if you are registered under .ir or the Iranian IDN ccTLD, that is very different than if you’re registered under an Arabic script gTLD because if you just see the Arab script gTLD, you may not know whether that’s Persian or whatever language that can be expressed by the Arabic script. Similarly, with Han characters, if you’re hancharacters.jp, you have an indication that’s probably Japanese. But if it’s hancharacters.asia, you lose that capability of knowing whether that’s Japanese, Korean, or Chinese.

So whether that was taken into consideration and how that was a consideration, it would be interesting.

My second question is actually building on that. There are a situation of mixed scripts, and I think you identified both Japanese and Korean – and in fact, in the cases of Chinese as well, if you call the Latin mixed script. How are they handled? I see Korean there, which I find interesting. When you talked about mixed scripts, does that mix the Han script as well? Especially on those few, where there are mixed, including Chinese, of the 29 tables, are we talking about 29 physical tables? Or are there situations where languages have multiple tables that eventually had to be combined by the implementer?

I'm sorry for the long comments and questions, but I think they are important questions. So hopefully you got it.

SARMAD HUSSAIN: Thank you, Edmon. Michel, you want to respond to that?

MICHEL SUIGNARD: [inaudible] one where we have both Latin and Cyrillic, if I can answer that. Otherwise, for mixed scripts, we added Latin mostly in the Asian scripts. That means Japanese and Chinese and Korean. We didn't put Hanja in the Korean table yet because there's no precedent of having Hanja in second level domains so

far. When we see a good reference for it, we'd be happy to add it, possibly.

To some degree, for the language [inaudible] I was thinking of the feedback I got before. We used CLDR. CLDR was a very fundamental reference for us because it's used, in fact, in quite a bit of platforms to define what is the coverage for a given writing system.

I understand that language and script are different animals. To some degree, when people want to create a new domain name, they may want to be just script-based, or they may want to be language-based. I think apparently at least ICANN is seeing a need for language-based definition for reference tables. It doesn't mean that you can't have at the same time script-based tables. I think one does not preclude the other one to exist, so if you want to create a script-based table, you can do that.

But you can do both of them, and there's nothing wrong with that, because people may want some time just to make sure that this domain name makes sense for a given language. So they basically intersect. I don't see them as being exclusive of each other.

SARMAD HUSSAIN: There's one more person. If you have a quick comment back, let's take that.

EDMON CHUNG: Follow up because one part that you didn't cover is the context of gTLDs versus ccTLDs. Because of the script situation for Han, for example, a Han-exclusive string.jp you would know is Japanese. But a Hanstring.gtld? You don't know whether it's Chinese or Japanese. So that might fail in terms of their LGRs. Was that taken into consideration in creating it?

MICHEL SUIGNARD: So far, we did a Japanese LGR and we did a Chinese LGR. The Chinese is based on .asia – zh, in fact. That's what we use. But we created it in the way that we created – basically would do a script-based LGR that you can use for [inaudible] traditional simplified [inaudible]. There's no bias in the table. It can be used as a way so you can get all the traditional simplified variants on [inaudible] original label, so we preserve even if it is a mixed script, for example, a mixed traditional simplified.

So we basically developed a Chinese LGR that can be used we see for both traditional and simplified, but it's meant to be used for Chinese, not for Japanese. There's a different user set size. The Chinese LGR is 19,000 and some change, whereas the

Japanese LGR is a typical JZO subset, which is about 6,300 and some change. So they're different in size.

You could do a unified one, but then you would have the variant set that will get in your way because if you create a Japanese LGR, you don't want to have to deal with variants unless you have to, frankly, whereas we see in China that you would need to deal with the variant set. So that's how we did it.

SARMAD HUSSAIN: Thank you. Over to [inaudible]. Let's take a last question or a short comment if you'd like, please.

UNIDENTIFIED MALE: Okay. Thank you. [inaudible] from the Arabic GP Task Force, IDN TF, and from Sudan Internet Society, .sd, the registry for Sudan. I just have a comment and a couple of questions. I don't know if I will be in a hurry. No, probably.

First I want to talk about mixing characters from different languages in the same script. As Michel said, an example from Iraq, if you want, maybe some company has in their brand name some characters – it is acceptable and how will it be developed differently in the root zone? The second LGR, at which level is it acceptable or not?

The second thing I want to talk about is Ajami. Yeah, as you talked about, it's very difficult for African language. I am from Sudan, and you have many African languages there. I visited an African university. It is a big university of [inaudible], where you can find people from all around Africa using different languages [inaudible] in both scripts, Latin and Arabic.

For example, they even have printed books using some characters which are not there even in the Unicode. They are using just old-fashioned printers to bring these books. So there's not really standards. It was very difficult for us and for the script-level task force IDN, maybe both in the Latin. Maybe they will face the same problem, but they making more standards and we are trying to help in that. This is my comment on Ajami.

Also, another last question for the IDN Guidelines Working group about the reports. What kind of reports will we expect it will be? RFC reports for the GPs and for whom is it targeted to use that report? Thank you very much.

SARMAD HUSSAIN: Edmon, you [inaudible]?

EDMON CHUNG: Yeah. Quickly, on the IDN implementation guidelines, the final product is focused on registries and registrars for gTLDs, and then also for IDN ccTLD registries. So that's I think the target.

However, when we produce the initial and interim report and the final report, we're trying to get the entire community to weigh in. As I mentioned, there are implications on contracts, on how IDNS are actually being used.

SARMAD HUSSAIN: If there are any more discussion, we can take it during the break. Thank you very much. We're already over time. Thank you for coming and attending this session. Now I will close this session. Thank you.

[END OF TRANSCRIPTION]