SINGAPORE – IDN Root Zone LGR Public Workshop - Integration & Generation Panels
Wednesday, March 26th 2014 – 13:00 to 17:00
ICANN – Singapore, Singapore

NAELA SARRAS: Okay, let's have people take places, please. We have plenty of room at the tables if you want to join us. Okay, so we're going to try and start in a couple of minutes.

Okay, I'm going to go ahead and get started, please. Sorry for the delay. We had another meeting in this room and we were back to back. Since we're a small group here, I'd like to encourage people to come up to the tables and sit with us. What we'll do is why don't we take a look at the agenda that we have planned for today.

This is something new we're trying, and I'm extremely interested in feedback on how this goes. For the people that attended this morning's session, we talked about a lot of work that's going on within the IDN Variant Program to implement the IDN LGF (Label Generation Rules) procedure. So we thought this might be a good chance for people that are coming to the ICANN meeting to maybe talk about different aspects of the procedure.

According to the agenda here, we've dedicated the top part to talking about the procedure, the community's role in creating panels, and so we'll give a little presentation about that, and then we want to hear from the communities.

So we have planned a presentation from the Arabic community, and then some panels that are at various stages of formation. Chinese, Japanese, and Korean panels will give us an idea of where they're at, and then the [inaudible]. At least today we should hear from Chinese and Korean. I'm not sure that we're going to hear from Japanese, but we'll deal with that when we get there.

Then we want to talk a little about the Maximal Starting Repertoire, which is out for public comment, and then we will finish with what we're calling here training on the XML. This is the specification being developed for a standard to represent Label Generation Rules in XML format.

So without any further delay, I'm going to go ahead and ask Marc Blanchet to take us through the first section of talking about the community work to form Generation Panels.

MARC BLANCHET: Good afternoon. So this part of the session is to describe how to form a Generation Panel. As the ICANN CEO said yesterday during the cocktail, I think Generation Panel is one of the very important part of this whole process because it's the community telling what should be supported in the root zone for their own languages and script. And as he said, I think we should be all there to help the community to create those panels and help them to work on it through it.

So that should not be necessarily news for most of the people here, but essentially the process that was agreed upon was a two-level process where Generation Panels that are created per script or writing system

actually do their work in terms of defining the code points and the variants that should be for their script. Then they send their proposal to the Integration Panel. The Integration Panel is the five people that are in this room and were presented this morning. The role of the Integration Panel is to assure that the proposal is fine and across all the scripts and so we have a [accordance] in the root zone.

The way the processor is written is that the way the Integration Panel works is they can only reject or accept the whole proposal. We cannot say, for example, "I don't like this, or this, or this." It's one or zero – acceptance or rejection.

And then the Integration Panel takes all those proposals and the GRs and then merge them into a unified LGR, which becomes the standard for the root zone.

As you can see from the previous slide, the root of this work is to actually have Generation Panels. ICANN issued a call for Generation Panels some month ago. The Generation Panels are the focal point for a given script or script family, and it's all about the community within that script or script family. It should be based by the community, driven and established.

Members of a Generation Panel should represent not only the languages or the regions involved in the script, but also on the other side of this equation is also expertise – so expertise, obviously, in the linguistics of those languages, policy, Unicode, IDNA/DNS, registry, and registrar. However, if a Generation Panel may not have all those expertise, they may be complemented with support from ICANN. The

communities should secure funding for its work, such as meetings, and when the Generation Panel seems to be ready, then they submit their proposal to ICANN to form a panel.

So right now, this is a very important time to get involved because this process is ongoing. The way you do it is to go the project's workspace, which is actually listed at the end of this presentation – the URL – and then you get the documents which tell you what you need to do to form the panel and volunteers and criteria and expertise needed and all that stuff. You should contact ICANN staff for information and also to help you in this process.

One thing which is happening right now, as we discussed this morning, is the fact that the MSR (Maximal Starting Repertoire) is currently in public review, so this is the set of code points that are acceptable to be reviewed and chosen by the Generation Panel. So this is very important that you review because for your own language, making sure the code points you need are actually acceptable in the set of non-excluded code points within the MSR. So this is very important for people here to understand that you need to review them so if there is some excluded code points there, then you should signal to us.

Later on, when we'll receive Label Generational Proposals, then there will be a public review, so that would be a time for you to review the LGRs of others, and then also the integrated LGR when the Integration Panel will have done its work.

Which panels are needed? The repertoire version one covers eighteen applied-for scripts. Additional four scripts are related and need to be

considered simultaneously. This is the decision of the Integration Panel: to make sure that we have [inaudible] and we take the comprehensive view of all the scripts that are related. So we need a Generation Panel for all those scripts.

The current situation right now is that there are some scripts that we haven't heard activity or intent or input so far, which are listed here. There are some scripts that have been more active up to a fully-formed Generation Panel. Obviously, again, this is a call for having the community to actually be involved, especially on the scripts we haven't heard yet.

This is an example, and it's just for illustration purposes. Sometimes there are related scripts, and the view of the Integration Panel is that we need to get a full view of all those scripts together. For example, here there's a Generation Panel L, G, and C, which doesn't mean anything. Those three are working by their own for their own script, but since they are related, they need to have some coordination and joint discussion. The Integration Panel will receive each of the proposals separately, but we expect that the Generation Panels would have coordinated and discussed together because they are related.

The way the joint discussion is done could be informal or formalized. It doesn't matter too much in terms of the Integration Panel, but we need to really make sure that all these related scripts are covered appropriately.

There's interactions between the Integration Panel and the Generation Panel. It's codified in one section of the procedure. So the first task at-

hand is to review the MSR and comments by the Generation Panel, or if the Generation Panel is not yet formed, at least the volunteers involved in this language or script. We have to recall that when the MSR is agreed, then it's become fixed, and therefore it's the starting point for the work of the Generation Panel.

We encourage Generation Panels to engage early with the Integration Panel so that we're making sure that all the potential points or problems are managed early in the process instead of later – not that we think that there will be some problems, we're just trying to work together to have this project successful.

These discussions could happen in different ways, and we're pretty open to any kind of appropriate way to have those discussions.

Again, the Generation Panels that have related scripts must coordinate between themselves, and again, recall that the Integration Panel's decisions must be unanimous, which means that any one of the five people have to agree. We cannot reject part of modify proposals. It's an all for nothing process. So, Naela?

NAELA SARRAS: Yeah. Thank you. Thank you very much, Marc. As I said in the beginning, we wanted this first section to be more about what the community is doing, so this is a little bit of what we expect to happen. Now we want to turn around and hear about what's really happening at the community levels to establish those panels.

Before we switch to the experience with the Arabic Generation Panel, which is a seated panel and currently doing work, I wanted to encourage everyone to please make this is as interactive as possible. Please don't let us just present slides. I guess I will speak on behalf of other speakers as well. If you have questions, please let us know. If you have any questions about the slides that Marc just spoke to, maybe we should talk about that first and then go to the Arabic panel. Okay.

Noha, did you have a question?

NOHA FATHY: Yes, I have a question. There is something that I'm not sure that I can understand clearly that was said in this presentation about the Generation Panels that have to coordinate our work together. As far as I understood, every Generation Panel would be concerned with just one script. And then you mentioned that there will be related scripts that might be sometimes of cross-cutting more than one Generation Panel. Could you please just elaborate more on this point?

MARC BLANCHET: It depends on the script we're talking about. There are scripts that are not related to others, so the Arabic panel doesn't need to talk to the Chinese panel, for example. So it all depends on which script we're talking about. It's only the related scripts that have commonality in terms of code points or glyphs that needs to talk together.

NAELA SARRAS: Thanks, Noha. Sarmad?

SARMAD HUSSAIN: I would like to make two points. I've raised these earlier as well but on this platform as well. First, I think, there is a value based on our experience with the Arabic Generation Panel work so far. I think there is still value for Generation Panels which are not related to still interact with each other because there's some phenomena which are common. So there are many different scripts, for example, which are using combining marks. And even if they decide to use combining marks differently, there may still be value discussing between those Generation Panels or different, in a way, the idea of solutions or possibly these different panels are considering. So there should be some framework within this project to make that possible.

I think one possibility is to use the LGR list more actively, but I think that's something we need to keep reminding ourselves that we should not just focus only within our own work, but also at least listen into what's going on. In that sense, also, if you have it in a panel, we should try to broadcast that information outside as well, so other people can also listen in.

So that's the first point I wanted to suggest. The second was that, as a member of a community, not particularly Arabic script community, but generally community, I think it is a good idea to review the MSR freezing process, where I would very strongly suggest that we give opportunity to form the Generation Panel to go through the MSR before freezing that particular block of MSR rather than have an open public comment period because a single person may not have perspective of the entire script block because there may be characters there which are outside

the language that particular individual will have in his or her scope. But as a Generation Panel, one would assume that the complete expertise of the script block is available to the platform and they can more I guess coherently reflect and comment on that script block. So that's it. Thank you.

MARC BLANCHET: Naela, do you want me to –

NAELA SARRAS: [inaudible] I'm sorry, Marc, yeah; mention your name for the participants.

MARC BLANCHET: I agree with your two points, which is first there's nothing that prevents people to talk and to help each other. I responded based on the slide that says you must coordinate for related scripts.

The second is the point you make about the MSR comprehensive and the fact that – so I think we currently have a review period right now, and I think it's up to the Integration Panel. And we already had some discussion about this about what to do next, depending on the results of the review period. So I think that's a summary to see what we will do after.

NAELA SARRAS: Okay, thank you. If we don't have more discussion on the first part, let's go ahead and go to the first community presentation, and for that we're going to go to you, Dr. Sarmad, for the Arabic.

SARMAD HUSSAIN: Thank you, Naela. How much time I have for this?

NAELA SARRAS: We have a generous 20 to 25 minutes, I think. Should be good, yeah?

SARMAD HUSSAIN: Thank you. This is an overview of what's happening for the Arabic Script Generation Panel, and the title basically does not say Arabic Script Generation Panel but Task Force on Arabic Script IDNs, and that is because this is a community-led effort and the community organized itself into a task force, which is looking generally at all relevant issues. And one of the things the community is very specifically looking at is the Arabic Script Generation Panel, but the Generation Panel has an ICANN-specific hat, "task force" is a community hat, which we are also representing here.

The Arabic script community has been active on IDNs for a very long time. There is a lot of experience in the community doing work for more than a decade now, and what we built as part of the Generation Panel, a lot of work comes from those interactions. And so it's I guess important to recognize those efforts which have come before us and are feeding into this process.

So I leave the details for you to look through, so this is available in the slides. As far as the Arabic script work in concerned, it is at this time currently feeding directly into 18 delegated – of at least accepted – strings, most of which are actually delegated, which include Shabaka and Bazaar, which are the new gTLDs, and the others are the ccTLDs.

The work on Arabic script IDNs is limited to mostly three different Unicode ranges: 0600 to 06FF, 0750 through 075F, and 08080 through 888FF. So there's a significant number of code points which the Generation Panel has to review to decide what needs to go eventually into the LGR based on the linguistic needs of the community and obviously balancing that against the stability/security issues for the LGR.

There have been additional characters identified during the initial process. This – obviously – work is a follow-up on an initial study which was done through ICANN. It was called Arabics. So there were different script studies. There was one specifically on Arabic script. It was called Variant Issues Project. Basically as it's a continuation in that process, we are considering all the characters which were initially recommended in that study, and then we obviously go and review those.

Other characters which were initially recommended by the community also include zero-width non-joiner and zero-width joiner. So there's still part of the scope of the work which we are considering. Obviously some of these become more restrictive because of the LGR process, which eventually came in.

We actually face multi-layer issues in Arabic as it's a reasonably complex script to work with in that context because we actually have, as I shared

with you, a large code point set, which has historical code points, but also code points that are very actively used. So as a first step, we need to basically figure out which code points are relevant for further work, but then as soon we start finalizing those code points, we run into a very interesting variant issues, which is going to be one the significant tests of our work as well in the future.

Here is some typology of Arabic variance. Arabic variance can come from various levels of various sources of types, so you have some letters which are totally identical visually but have different code points. You can have letters which are similar in shape and are confusable by linguistic communities. So they can still be variants even though they do not have exactly the same shape. Then you could actually have possible variants which are totally different in shape, but are perceived in some ways as still the same by the community.

As I said, we have to balance the community needs with the security and stability needs, as well as the Generation Panel is concerned, and we will obviously be taking care of that as we go forward.

The formation process of the community – I think this is probably more interesting in this context as well – was that there was an initiative started in the Middle East and adjoining countries which are using Arabic script in general, where we wanted to promote our DNS-related work in the area – not only technical work, but also capacity-building work business in this area. So we actually formed the Middle East Strategy Working Group, which is actively working towards promoting Arabic domain name system (DNS) and associated work in that region.

One of the things which the Middle East Strategy Working Group, which is a community-based group, identified in its very early stages was that there is significant amount of work which is needed on Arabic IDNs. So a task force was formulated to look into those issues, and that's who we are now: Task Force for Arabic IDNs.

UNIDENTIFIED FEMALE: [Inaudible] switch the slide on the screen.

SARMAD HUSSAIN: Oh, sorry. As far as the task force is concerned, we are actually looking at Arabic IDNs as a holistic issue and looking at holistic solutions. One of the first things which we recognized in the process was that we do need to work with ICANN and develop a LGR so we can address all of the variant issues which are faced by the community.

One of the first tasks we've taken up as a task force is to formulate a Generation Panel with ICANN, and we're actually working on taking that forward. But the other work is not going to be limited on the Generation Panel itself. We will also be looking at generally universal acceptance issues of Arabic IDNs, Arabic e-mail, and some of the other issues that the community is pointing out to us. So we're actually very deeply rooted inside the community and we're looking at the feedback we received from the community to drive our work and drive our work technical focus as well.

It's a community-driven process, as I shared with you, so we actually announced an open call for participation for Task Force for Arabic IDNs.

It was announced at the second Arab IGF in Algiers last year, and we have a standing call, meaning that it was not a deadline-based call. Anybody can apply at any time, and we'll consider membership to the task force, and it still is open if any of you are interested to join in.

What we do is we divide the task force into smaller groups. One of the groups obviously is working on Arabic Generation Panel, but we're very soon going to form another group, which will start working on universal acceptability of Arabic IDNs and so on and so forth. So it's a process in which we want to keep the community engaged. Wherever we go, we spread the message and try to invite more people on board.

It's a very, very open process. All the deliberations which we do on Generation Panels are through e-mails – most of them – and all the e-mails are publically listed, so anybody can go and review what was discussed, what were the opinions of people, and eventually what was decided. It's very publically open.

We also do regular calls. All the calls are recorded and are also available on our Wiki site. So it's a very transparent process where everybody is invited to join in, and even if you do not want to join in, you are very free to come and listen and see what's going on.

We actually gave the GP proposal to ICANN in January, but before we gave that proposal, we actually shared informally with the Integration Panel and ICANN in December. So that was a very useful process for us. We actually developed the proposal, sent it into ICANN and said, "This is what we're thinking. Is there any feedback? Let us know." They were very quick and they came back to us within a week that comments that

"This information may be missing. You may want to add those." So that was a very useful process, and I strongly encourage others to follow that as well. Then we formally submitted in January and the proposal was accepted in February, so we've been officially a Generation Panel since then.

We have 21 members on the panel at this time. There are still five or six pending applications and we're receiving new ones, so it's a growing panel. We come from eleven different countries, speak nine different languages as first languages of the panel members, but beyond that, many of us also have expertise in use of Arabic script in many other languages. And we come from a variety of disciplines: academia, both from linguistics and technical sides, registries, registrars, national and regional policy bodies, like League of Arab States and UNESCO, and also community-based organizations. We have members who are actually linguists working for a member of a community, which is our language [inaudible] of working to document their languages, so part of the very community-based organizations.

This is just details about where our information is available. I've already gone through this. If you are more interested, please come and visit us on our Wiki space and come and see through our e-mails and get involved if you are interested.

As far as our work structure is concerned, basically we did an initial planning. In our proposal, roughly we have three stages of our work. In the first stage, we do the code point analysis. In the second stage, we do a variant analysis, and in the third stage we do a whole label analysis, and each one is built on obviously the previous stage. Roughly we're

planning about three to four months for each of these stages, so we're talking about a year of work from end to finish. We started to January and aim to finish in about December this year.

But it is a lot of work for us, as I think a lot of members of the panel are now realizing after a three-day face-to-face meeting in Singapore right before this meeting. We have significant support from ICANN to do this through their global stakeholders outreach program.

For each of these stages, what we've done for code point review, variant review, and whole label review, we've actually subdivided our work in another three-layered structure. So for each of these stages, we go through three stages within the Generation Panel. First, we do not go into the actual analysis, but we do set up principles to guide us for that stage. For code point level, we first did an exercise of developing principles, which will help us guide which code points should be included or excluded.

Then once the principles are reasonably stable, we go into the analysis phase, where we're using the principles to analyze the data. Then after analysis, we actually do a public review call for each of those stages. So that's what the plan is. We actually do a public comment internally, not through ICANN, but through our own community and the larger community through ICANN mailing list, separately for code points, for variants, and eventually for whole label rules, and those individual code points are not part of the ICANN formal process. Those are internal community processes which we said we want to keep it very open and want to keep the community engaged in the process.

Eventually, when everything is going to be done and we package it and hand it over to ICANN, ICANN will also do a complete community-based process, but we do not want any surprises at that late stage in the game, so we really do want community and everybody involved with us right from the start.

Our principles document is already out. All this information is going to remain available on our Wiki as it gets produced, so feel free to come and look at our principles document. Feel free to give us feedback. We had a face-to-face meeting in Singapore just before this meeting, very useful. We've actually now progressed to the code point level analysis, and we aim to finish that in the April timeframe, just in time for the public comment period for MSR.

As I said, we are going to do a character level analysis and then variant level analysis and then whole label analysis, and that's sort of on a rough timeline when anticipate finishing this work. This I already talked to you about.

We've been engaging with the community as well. There was a Middle East DNS forum in Dubai early this year. We went back to them and discussed many of the technical issues, and there a very strong feedback from the community to start work on Arabic script IDN universal acceptance and Arabic script e-mail. So that's additional work which we actually will be starting very soon.

Obviously, as I said, we do have a Wiki space that is hosted on ICANN Wiki space. We are getting significant support from ICANN outreach program. For this, they give us support for our cause and also a face-to-

face meeting. I think, as a community, it's worth recognizing ICANN's support in that context. That's it. Thank you very much. I will take any questions. There are many more Arabic script [inaudible] IDN members here, and I'd request them to just very quickly raise their hands so that you know who we are in this room. We are going to be available for questions. We'll take some now and we can take some later. Thank you. Back to you, Naela.

NAELA SARRAS: Thank you so much. Maybe while people are thinking of their questions, I have more comments and questions. Really, I want to commend the panel and congratulate you and this tremendous amount of work you're doing. It's very obvious that a lot goes into this work, and I just want to thank you and congratulate you on that, and really do make it a point to say you are chartering the path and running into issues and trying it out for the first time, so this experience is really valuable for us and hopefully for other communities as they go through and build their panels; so just a sincere thank you to everyone here.

Professor Lee, go ahead.

DONG-MAN LEE: Yeah. First of all, my name is Dong-Man Lee from Korea. Thank you very much for your presentation. That really inspired me where to go, how we actually conduct our own language Generation Panel. So it's very insightful and encouraging. I think simply speaking, though we can just copy what the Arabic did. So his job looks pretty easy.

My first question is, well, you mentioned the LGR panel members are still growing, so I ask actually ICANN or the IG, is it okay? So it's not limited? The member of the membership is not limited? So we can actually continuously recruiting based on the need and recast?

NAELA SARRAS: If I may comment on that, that is correct. When the panel is first formed, we are looking at completeness of the panel in terms of the different expertise that's required to join the panel. That's looked at before the panel is seated. But the process also allows for additional members as the effort becomes more known in the community and more experts are able to join, yes, the process allows additional people to be added.

This question of the panel, and then the panel basically sends a request saying, "We gave you a proposal saying these 20 people, let's say, were in our proposal, but now we wish to add 2 more individuals," and then give the relevant information about these individuals.

MARC BLANCHET: Well, let me just kind of quickly check the process because when we first formed the Generation Panel, I asked all the potential members to send e-mails, and I'm not sure whether they got any response directly.

For example, myself, I made a request and I didn't get anything, so I just simply presumed the no response means yes, so we just kind of continued to. So next, if and when we actually recruit the new members and then we can – the condition is we have to always get the approval

from you. I just want to make sure whether [inaudible] or simple just part of the process.

NAELA SARRAS: Sarmad Hussain, did you want to – so I will go and go back to you? Okay.

I know which e-mail you're talking about. You sent a number of individuals. So the process is that we asked, are they also going to submit their own SOIs individually, and I don't believe we received the next one.

MARC BLANCHET: [inaudible] Some of them – I'm not sure how many people, actually see the reply [inaudible].

NAELA SARRAS: Speak – we have remote people. Go ahead.

MARC BLANCHET: Sorry. I don't know how many people actually have gotten response from you. For example, me? I haven't gotten anything. I sent e-mail. I just sent my short bio and so on, and I didn't receive anything.

NAELA SARRAS: I will personally look into that. That does not sound – so – I will look into that. Last comment before we go to Dr. Sarmad: the procedure isn't very prescriptive. There isn't one way to form a panel. In the case of

ICANN – maybe Dr. Sarmad can better touch on this – in the case of their panel, they considered the group and submitted as a panel. In their proposal, they submitted the group, so there weren't individual SOIs. The whole proposal came together listing who the members are.

So the procedure allows for both. It allows for individuals to come through as interested individuals, or it allows existing working groups to come through as proposals. So I'm not sure that there is one size fits all yet for this. Does that make sense? But I will look into, with my colleagues, why you don't believe you got the proper answer.

MARC BLANCHET: I certainly understand you tried to improve the integrity of the panel, but, yes, when you had the private meeting with the Integration Panel, we were advised to embrace more people in various areas, and if it's just a simple a matter of just checking, it's fine. But if it's like sending the request and then getting the approval, I don't quite understand why that is required.

NAELA SARRAS: Okay, let's [inaudible] it's possible that maybe what you have to add might shed better light on this, and then we'll come back to you but maybe you might –

SARMAD HUSSAIN: I just wanted to narrate a little more about the proposal process from Arabic script GP, and I think that may, in some ways, contribute to this discussion.

When we were applying, we actually had both choices, the two ways of doing this, at least two ways of doing this. One is that all individuals directly apply to the ICANN mailing list by mail or e-mail, and ICANN tries to formulate that. But we found it more convenient since it was a community-based effort to formulate the panel ourselves outside ICANN first and made sure that it meets the requirements which have been stipulated in the LGR process, as far as expertise in diversity and languages and so on is concerned. And then apply and block together to ICANN as a single proposal rather than everybody trying to do this individually.

In the process, what we did request from ICANN was that those who had applied directly to ICANN and not somehow learned through the community process, they check with them and if it is convenient with them, if it's okay with them, ICANN can forward their request to us as well so that we include them in the process so that everybody eventually who's either going directly to ICANN or coming to our community process, we all still formulate a single group. And since it's an open panel, everybody can join in, so there's no restriction on the website.

We eventually put everything together and handed it over to ICANN, and since then we've also been asking ICANN that we need to add more people and we were now formulated a concrete process for that, and it's working out, so I think we've really not found any challenges in going through this process so far.

Our proposal, if you are interested in looking at it, even though the process does not require it, but we've still posted it for public

information and it is available on our Wiki site. So if you're interested in looking at our proposal, it's out there as well.

NAELA SARRAS: Thank you so much. That was very helpful. Yes, Asmus?

ASMUS FREYTAG: I was wondering maybe whether you'd want to say something about what the main interests are and even looking at the composition; you mentioned one which is in the beginning to make sure that there's the minimum required distribution of competency, but the other thing that is described in the procedure document is that there is theoretically the possibility that the Generation Panel gets captured by one single interest, and by them just joining in very large numbers. That is one of the things that you are looking for when the panel is growing later, so that is one of the reasons why you are interested in finding out who is joining and stuff. It's not just to create bureaucracy. That is [verily] important to –

NAELA SARRAS: Did you want to comment on that, Professor Lee? No? Okay. Okay. Okay. I'm sorry? Yes, we should. We will have this discussion afterwards.

Are there other maybe questions or comments on the Arabic? Okay. So why don't we move ahead and I think we had planned at least another – actually, several other interventions – so [inaudible], can we call on you next? We grouped here that we would like to see maybe an update on

some of the panels that are in stages of being formed, and I think Chinese is a fairly advanced one in that stage. So would you like to give us a little bit of an update with what's going on there, please? Into the microphone please and please state the name, even though I'm not –

UNIDENTIFIED MALE: Hello, everyone. My name is [inaudible] from Beijing. To begin with, what I'm talking about how I would like to present the background and where we are. Okay.

Chinese characters are logograms used in writing of Chinese and some other Asian languages. They're called hanzi in Chinese, kanji in Japanese, and hanja in Korean.

Because Chinese, Japanese, and Korean scripts share common background of Chinese language, the common and shared characters were identified and then the [CGK] unified ideographs in the process of Han unification. Chinese hanzi, Japanese kanji, and Korean hanja often referred to as ideograph, which is a graphic symbol that represents an idea [inaudible] Chinese, Chinese hanzi, Japanese kanji, and the Korean hanja have been merging [inaudible] unified ideograph and that extension in Unicode. Okay, that's the background.

Last week, a meeting was held in Beijing to initiate the Chinese Generation Panel. Hansang Lee as the observer from ICANN was there to get involved in our discussion. During the meeting, the Chair Elect, who by the way is Dr. [Wei Mao] from CNNIC and by the way, the member of CGP was pre-determined via mailing list and approved by ICANN, okay?

Besides, the meeting further discussed upon the joint CGP-JGP [PGP] and the Integration Panel, the potential collisions might happen between them.

All the discussions can be concluded as follows. [inaudible] between Integration Panel, Generation Panel – maybe we can call it CJK community, should be well-established to solve the potential collisions caused by overlapping of character strengths.

The second, Chinese, though used to be employed by Mongolia, South Korea, [inaudible] long losing its practical demand. That's the Integration Panel work is confined within China, Japan, and Korea, and [inaudible] other ways Integration Panel should be met in order to seek understanding and reach consensus. Third, focus on the milestones and make it flexible. Four, we're formulating character table [inaudible] a second column and the reduced quantity preferred variant if necessary.

The last important [inaudible] CNNIC team is responsible for the comparison of the character table between MSI and the current CDNC to make sure when the CDNC character table is covered by, if they are. If they are, the CCNIC shall offer some proposal for CDNC. By the way, CDNC is short for Chinese Domain Name Consortium. Okay, that's all about the meeting.

On the Chinese Generation Panel, we is honored and delighted to know that its contribution so far to ICANN's IDN ccTLD process have proved fruitful and helpful. [inaudible] contribution at work, the CGP is in the process of implementing an action plan based on its strategy document in order to perform tactical training and dissemination activities.

So I'm taking [inaudible] approach on current technical matter including the following: a Chinese script at a level generation root set for their own root zone. Second level, LGRs for the Chinese script, Chinese script internationalized registration data protocol and practice, universal acceptability of Chinese-grouped ideas and their variants, technical changes around registration of Chinese IDNs and variants, operational software for Chinese script IDN registry and registrar approaches, and a DNS security matters specifically related to Chinese IDNs and variants, technical training material around Chinese script IDNs.

Okay, the Generation Panel intends to divide our work on LGR for the root zone into a fourth stage. They're organized as follows: Formulation of a code of points, formulation of variants, finalization of whole label rules, and the last one would be a LGR document for Chinese group for ICANN.

Okay, so all the work will be done before the ICANN meeting held at Los Angeles. The rough milestones would be formation of a Generation Panel by the end of March this year, which, by the way has been done, to get the character set by the end of May, and to have a definition of general principles of variants by the end of July, to have the label rules in early September, and then our panel will be finalizing the LGR document before the Los Angeles ICANN meeting.

So I guess that's what I would say. Of course, the panel, we've already rededicated ourselves for the important work for there to come. Thank you. Chris Dillon, do you have something to add?

CHRIS DILLON: No. I think you've said it all very neatly. Thank you.

NAELA SARRAS: Thank you very much, [inaudible]. For the benefit of everyone in the room, we didn't prepare slides, so it was just a talking presentation. But I'm really interested in the milestone timeline. That seems like a lot of work ahead of you, so –

UNIDENTIFED MALE: Yes indeed.

NAELA SARRAS: So, good. And thank you. Any comments for [inaudible] on the talking points? Go ahead, Han Chuan.

HAN CHUAN LEE: Hi [inaudible] to thank and commend the CDNC for supporting this work and for convening the CGP and starting work and, as you can hear from [inaudible], they have done quite a lot of work, and I wish you all the, the CBP all the success in getting the LGR. Thank you.

NAELA SARRAS: Indeed. Thanks, Han Chuan. Tremendous work, there. work there, Thank you. Go ahead, Dr. Sarmad. Go ahead.

SARMAD HUSSAIN: Just a question on how, from what I understand, there are three different panels on Chinese, Japanese, and Korean. If that is the case, I'm just very curious to learn how the three panels are going to be coordinating with each other, and also how we can coordinate with you on work you're doing and work we're doing. How is that going to be possible?

UNIDENTIFIED MALE: Okay, thank you. Very good question. Actually, the Chinese Generation Panel considered to establish an inter-Generation Panel to coordinate. And actually, the Chinese Generation Panel had a face-to-face meeting with the Integration Panel, but it's now this.

What we want is maybe ICANN crew to establish an inter-panel community to coordinate this. We don't care what it's called, as long as the coordination mechanism could be well-founded. So maybe we can coordinate CJK community or CJK consortium, as long it's authorized by ICANN and approved by ICANN.

NAELA SARRAS: Let's explore two things that are going on, I think. The first one it I think Sarmad was alluding to and it's come up now is we need as panels to come online, and even panels in the formation stages, we need some coordination and cross-fertilization of information. So if one panel ran into something, another panel doesn't have to go down the same road. We can learn from each other.

So for that, what we had envisioned as staff is to create a mailing list called LGR@ICANN.org. Our original thinking was, "All right, as the Integration Panel is of course subscribed to it, and as panels are added, we will subscribe them to the list." So right now we have pretty much the Arabic panel and Integration Panel on that list.

But help us think through if we didn't have the logic right because until the panel is formed, at least, at the current rate, the way the process is set up is until a panel is formed, they won't be added to the list. So tell us. Let's talk a little bit more about what – of course we're there to facilitate, so if we need to create more lists or repurpose lists, we're willing to do that if it serves the community. And then we'll come back to you – to what you're talking about, the point is that the coordination or whatever the label is, but let's talk first about the coordination among the bigger picture panels. Does that make sense?

UNIDENTIFIED MALE: Yeah. I think it's [inaudible] on discussion. But I would be very happy to hear some suggestions from Professor [inaudible]

NAELA SARRAS: Okay, so it sounds like we want to talk about that issue first. So the other issue, [inaudible] you're raising is coordination amongst Chinese, Japanese, and Korean panels, right?

UNIDENTIFIED MALE: Right.

NAELA SARRAS: This is more of a community thing, really. It's really up to the community to discuss this and decide the model for them, and then communicate it to ICANN. Maybe what we can do is first have somebody from the Integration Panel comment on how the procedure approaches this coordination. And then we'll come to I think I have Akshat and who do I have? And Sarmad.

ASMUS FREYTAG: Addressing this from the perspective of what the document says that's entitled "Procedure for the Development of a Root Zone LGR with Respect to IDN Labels," which is really, our, if you may think of it, is our bylaws for this operation. This is our founding document for the whole effort.

The procedure, if you read it, leaves two things open. It leaves open the number of scripts a Generation Panel can work on simultaneously, which is in the case for the Indic Panel, there's an interest to do a single panel for multiple scripts.

The second thing, while there is a requirement in the panel that related scripts of overlapping LGRs are well-coordinated, it is clear that it leaves the organization of the work of the Generation Panels to the Generation Panels. So other than the requirements of how and when to make a submission and how and when to communicate the membership to ICANN, there are no requirements on how you meet and how you organize, whether you form subcommittees, whether you all do it in one session, whether you do it by mail, by phone, in person. All of those

things are on purpose in the procedure, and Sarmad can remind me as well because he was also part of the effort of writing the procedure, to left that open to allow the community the essential flexibility to be able to carry out the work in the best manner possible.

So from the point of the procedure, the important thing is what is the end product of this effort? The end product of this effort should be a series of coordinated LGRs, in the particular case of CJK (the Chinese, the Japanese, the Korean) that do not have a conflict on a technical level, that are agreed to by the communities, and that are submitted to the Integration Panel for integration and review. That is what we're looking for, everything else is really –

We have members from the prospective Korean panel here. We have members from the prospective Chinese panel here. Unfortunately, I don't know whether we have anybody from the prospective Japanese panel or anybody who has knowledge of what they would like here, which would make it easier. You two or three need to get together and come up with some model that works for your community.

UNIDENTIFED MALE: Thank you.

ASMUS FREYTAG: I don't know whether staff even has the ability to endorse anything else because it is not covered by the procedure. It's not forbidden necessarily, but it's not covered by it.

NAELA SARRAS: Thanks, Asmus. I believe Sarmad was first. So let's go to you, Sarmad.

SARMAD HUSSAIN: What's probably called for here is some facilitation through the ICANN team because in many cases, these panels are reasonably independent within community. They have their own outreach, but a cross-community, sometimes that does require a facilitation role, I think, and how that's exactly done is something that you eventually I guess we'll figure out. But I think that probably just needs to be highlighted. Thank you.

NAELA SARRAS: Thank you. Akshat?

AKSHAT JOSHI: Hi. This is Akshat from proposed Neo-Brahmi [inaudible] Panel and I had some of the comments about how we can manage from ICANN's side the communication between the panels and the way they have been designed in the procedures. If I recall correctly, there are two kinds of communication that are there in the procedure. One is the Generation Panels interact with the Integration Panel, and they coordinate among themselves as well.

If we go back to the first phase of the project wherein we had separate case studies, I remember ICANN preparing separate mailing lists for each of the case studies, and then there was a generic mailing list, which was VIP, to which everybody was subscribed, which in this case ICANN sees as LGR@ICANN.org to be that case. Maybe when each

Generation Panel is formed, each of the panels can be given a specific e-mailing list in addition to that, like instead of keeping the dialogue with the Integration Panel and the Generation Panel to the LGR mailing list, which is generic. Maybe we can have a separate mailing list only for the Integration Panel because those queries that say, for example, Neo-Brahmi [inaudible] Generation Panel may help the Integration Panel. It may not be of interest to every other Generation Panel, so just to separate out that, I think that could be one of the projects.

ASMUS FREYTAG: I have a suggestion, a practical one along that line. I think it's an interesting idea to have these focused mailing lists. It gets a little difficult to manage, but we don't have that many Generation Panels active at the same time.

What might be of interest is to give the chairs of every single Generation Panel a courtesy observer status. Give Sarmad, Chair as the Arabic Generation Panel, an observer status on, say, the Chinese, too, Integration Panel traffic so that if issues come up that are of generic interest, he can note them and say, "Okay, this is what we're interested in," and then approach the other panel on a different channel and follow up. I think that would aid the transparency without creating super traffic because I think Chair-people will be very disciplined in not abusing that privilege.

NAELA SARRAS: Just so I understand the suggestion correctly: currently there is a seated Arabic Panel and they have a mailing list. Is it coordinated through ICANN? It is, right?

SARMAD HUSSAIN: We're not using the one coordinated through ICANN. We have of our own community an open e-list, which we do.

NAELA SARRAS: There's your answer.

ASMUS FREYTAG: Okay. No, I thought the idea was that there was going to be a list that one panel can use to track its issues with Integration Channel. Was that your suggestion? Or it's internal?

AKSHAT JOSHI: My conversation was we just create one mailing list to which only Integration Panel members would be subscribed to, so that whenever we want to talk to the Integration Panel about some specific cases – similarly, there will be every Generation Panel wanting to talk to the Integration Panel. So instead of populating all those things on the single mailing list of LGR to which everybody's subscribed to, only the Integration Panel gets those mails. That can be public. That's not an issue.

But then only the Integration Panel from that point can look only at that mailing list to see what panels are communicating to them.

ASMSUS FREYTAG: Yeah, I just wanted to not have a model where we have lists that one Generation Panel and Integration Panel is on a closed list. That I wanted to avoid.

NAELA SARRAS: I second that. Other lists we're proposing to have here are going to be open archive lists, for sure.

Okay, I really am happy to entertain this a little bit more because we dreamt up the solution of LGR@ICANN.org and if it's not working, we need to enhance it.

In the GP document of how to help set up the GP (Generation Panel), we did say that if a panel needs us to set up a panel discussion list, mailing list, we can do that. Not everybody needs it because case in point here, the Arabic Panel has their own. So that's no problem, Akshat. We can do that, but now we need to flush a little bit more on how to get interactions between Generation Panels and the Integration Panel more visible, right? I think that's your concern. And target it so that not everybody has to see everybody's issues, right?

AKSHAT JOSHI: Just try to segregate those two things. One is when we communicate with the Integration Panel the way currently it is being seen, there is

only one mailing list, LGR@ICANN.org, and if we have something to communicate, we'll be communicating on that. That is the way it is designed. Okay? So when I see if I'm on that mailing list and say, for example, Arabic Panel also has to communicate something to the Integration Panel, we will also be receiving those mails here, likewise for every panel will be receiving those mails.

So I think from, I don't know, from our side it will be like if they're not considered to be related, which is not the case with CJK, but then there are fairly identified separate groups which are separate. So from our side, as well, then it becomes manageable. If there is only one specific alias for the Integration Panel to which communication with forum and communication with the Integration Panel happens with that, and even for the purpose of inter-panel communication, if a panel has a specific mailing list, and if we want to cooperate with the Arabic Panel, we'll just send them mail across to their mailing list and what aliases could be in touch with each of that. That way, inter-panel communication can also be facilitated and I think that will be more manageable for each of the panels, I guess.

NICOLETA MUNTEANU: Hi. To clarify, we have this LGR mailing list for the Integration Panel. Members are subscribed and are able today to receive any requests from any community member.

We also subscribe to this list for the sake of transparency the members of the Forum Generation Panel, which is the Arabic one. However, you can still view that the Neo-Brahmi community members are working to

create this GP proposal can send to this mailing list and the Integration Panel can respond. In the meantime, the Arabic Generation Panel sees what is being sent. But you are not subscribed to it, which can change depending on what we decide. We can subscribe everyone and everyone receives everything.

On the other hand, we also have this Arabic GP mailing address, and we can create a mailing address for the Neo-Brahmi, and the two communities can communicate between them if they wish. The archives are public, but this will not go to the Integration Panel. This will go in community Generation Panel interaction. But we can also improve these tools, depending on how you request and depending on the experience I think.

NAELA SARRAS: Let's hear from Marc. It looks like he's sitting on a solution.

MARC BLANCHET: You will tell me after my comment. I'm kind of wondering if we're trying to do too many stuff for maybe one or two e-mails for the next six months. I just find with hundreds of mailing list, I'm questioning if this is really needed. That's it. But I'm all fine for all communications and speaking with everybody, but its –

NAELA SARRAS: What we're going to do is we're going to talk to you individually after this. Really get this down so we can address it correctly, here and with [inaudible].

I think what is wise to do at this point is everybody who expresses an interest in being on a Generation Panel should probably be considered a candidate to be added to the LGR@ICANN.org mailing list because I think that would solve a lot right now. So instead of creating many more lists, let's start by, if you've already expressed interest, let's add you to the LGR@ICANN.org, and we will communicate better on that list that there is a Wiki space that lists how to communicate to a panel. So for example, the Arabic alias, that exists already, so you know how to communicate to contact them, and we'll go from there.

I wanted to ask if we have anyone online. Any questions? Any comments?

Okay, good. Asmus, can we close that topic?

ASMUS FREYTAG: I was just going to violently second all these notions because I think that the mailing list design in committee, especially in an ad hoc community like this, is never going to go anywhere, and Marc I think has it absolutely right in saying, "Don't engineer a solution until you have enough traffic to divide it up." So don't make more buckets until one of them is overflowing. So just take it offline and do the design offline.

NAELA SARRAS: Good. That was very helpful. We are on the slide that sees Chinese, Japanese, and Korean panels. I believe Professor Lee is here with us. Could you give us – oh, we have a presentation. Yeah. Yeah, yeah, yeah.

Yeah, yes. We weren't aware of the slides, or I wasn't. Let's give the – oh yeah, give him the – excellent. No, I just wasn't aware.

DONG-MAN LEE: Okay. Good afternoon, everyone. My name is Dong-Man Lee. I'm from KAIST, Korea. I'm also a member of the Korean Internet Governance Forum. I've been working with the Korean Internet community the last 16 years.

Well, the last ICANN meeting, we were advised to form a Korean LGR panel as soon as possible. So I got that message and went back to Korea last November and started the formation of KLGR the members, and we've been successfully the recruiting people at the starting point, and we're going to expand to include more of the members for covering various aspects of the Korean language. Okay, so I'd like to share the current status of the Korean LGR with you. Your comments and the suggestions are very much appreciated.

Okay, at this point, there are the eleven members from various organizations with various expertise levels. As you see, I'm just playing the role of the Acting Chair. I don't claim myself the permanent Chair. Based on today's and yesterday's discussion, I think the code expert, the person who has the expertise on code needs to be Chair. So me, because I was asked at the first place, I thought the last time it was very urgent and important, so I took that responsibility, but too many things on my plate. So to make this thing successful and done, I think I'd better find the real Chair.

Anyway, at this point, we actually formed our panel – rather the technical people, but later yesterday we were advised to have non-technical people, such as the linguistics and culture and societal experts, because they use the Korean language and domain names, not only just the technology but also some cultural and societal issues as well. Yeah, we'll take that advice and we'll include more people.

Some of the panel members are probably well known, way back, ten years ago the JET CJK Collaboration, so Professor [inaudible] came and also Dr. [inaudible], who actually had played a great role in that time. They will play a major role in this realm.

Let me just kind of share quickly the Korean language specifics. As you know, Korean script has the hangul syllables. Of course, there's the H consonant and the vowels already in there, but we actually treat them, as far as cyberspace goes, we have combined the syllable. So 11,072 characters registered in Unicode space. That's the code point. We also, the register unified the Hanja characters range in the region, which is agreed by all three countries at the Unicode space.

As far as the local IDN, the domain names are concerned, we have we have .hanguk and even .kr. We allow the Korean domain names under .kr and .hanguk. But all domain names at the second level, they only compose by the Korean characters, English alphabet, and digits and not include the hanja – Chinese characters. So we have to make a decision on whether we're going to include the hanja at top-level Korean IDN names because which is not our local jurisdiction.

As far as hangul goes, well, we use the ideogram – the character set – so we do not have any variant issues, so it's pronounced as it shows. Of course, there's some dialects, so depending on which regions you live in, there could be different pronunciation of the same character, but only focused on the standard pronunciation, we don't have any variant issue, and so on. I've already explained this. One pronunciation, one syllable.

I've already explain all han characters, right, including Korean hanja defined for CJK Unicode, and we're more than happy to participate, and as we did in the JET's time.

But how to join? We are not fully aware of the current status of the CJ joint work. Last time, when I attended the November ICANN meeting, the China and Japanese people already worked together, so are more than happy to get updated on your current status, and then we'll try our best to speed up our learning curve to catch up your current status.

Finally, we have some of the historical difficulty, such as even though we use the same language, but the last over 60 years, we developed different words, meanings, but if this issue is only focused on the syntactical aspect of Korean language, I don't think it is a big problem. But there is still how we're going to incorporate the North Korean side if they later raise their hands on some disagreement on this, then it's some kind of issue. But at this point, frankly speaking, I don't have any clear idea how we're going to deal with this.

But still, as I said earlier, we're going to invite the linguistic experts and the social science experts and political science and ask about this issue

of how we're going to resolve this problem in a more safe and reasonable way.

The more imminent issue is whether we're going to include hanja or not. Current consensus at this point, as far as current KLGR members are concerned, the consensus is not include hanja, but, well, when we had the discussion with the IG yesterday, well, I think I would need to reconsider that part more seriously and have the communication with members and then make a decision soon.

As I've said in the previous slide, coordination with the CJ LGR team members and how we're going to collaborate on the CJK, unifying the han character LGR document generation. Okay. That's it. Thank you very much.

NAELA SARRAS: Thank you and thank you for your work on this. You're right. You quickly realized that this is going to be a lot of work. Are there any comments or question about Professor Lee's presentation? Rinalia, please.

RINALIA ABDUL RAHIM: Professor Lee, Rinalia from Malaysia. I'm on the Arabic script Generation Panel.

DONG-MAN LEE: You can call me Dong-Man.

RINALIA ABDUL RAHIM: Dong-Man.

DONG-MAN LEE: Yeah, I'm not here to teach, do you don't have to "professor" here. Thank you.

RINALIA ABDUL RAHIM: Okay. I didn't quite catch the end of your point in the last slide about the inclusion of hanja. The consensus was to include it or not to include?

DONG-MAN LEE: We haven't decided yet.

RINALIA ABDUL RAHIM: [inaudible]

DONG-MAN LEE: Okay. Right, right. But as I said earlier, as far as a Korean second-level domain names under the .kr, .hanguk, we only uses hangul, not including hanja. So the top-level gTLD, well, is a different issue, right? So we have to think about most carefully.

NAELA SARRAS: Okay. Regrettably we don't have any of our colleagues from the Japanese side to speak today, so unfortunately we won't have that part today. It would have been very helpful indeed. So unless I have more

comments of questions on these three panels, we can go ahead and go to the next presentation, which is going from Akshat from the Neo-Brahmi Generation Panel. Please go ahead, Akshat.

AKSHAT JOSHI: Good afternoon, everyone. I'm Akshat. I'm [inaudible] from C-DAC, which is the Center for Development of Advanced Computing. It's an autonomous body and there is a Department of Electronics and IT, Ministry of Communications and IT, Government of India. We are a body which is working mostly on the group, specifically in CDAC in which I represent, is working heavily on the nicer language processing side. The way were came into touch with internationalized domain names was through the .in registry, when they were going for IDN ccTLDs, and the government work that was done in that. To that, we became more aware of the issues and got involved in these IDN-related things.

We were part of the initial phases of the project as well, like the case study deemed the Integrated Issues Report, the procedure drafting DNS work, and now that we are into this phase, now we are gathering the community to participate along with us so that we undertake the task of bringing out the LGRs for this set of scripts.

This is a brief overview of my presentation. Firstly we'll be talking about what is Brahmi in the first case because initially we were only talking about the Devanagari script and what is Brahmi and what is this new thing. Why Brahmi and why not go the separate script away? We have different scripts and there is always this possibility that we could have separate Generation Panels for separate scripts. Then why the new

Brahmi approach? Then we'll be talking about the linguistic scenarios associated with the Brahmi-derived scripts and ongoing efforts from outside, and the challenges that we currently face.

Coming to the point of what is Brahmi, it's an ancient script that evolved during the final centuries B.C. It's an old script. It's not being used today, but it has acted as a metrical script for many of the modern scripts that are being used in the region today. There are many levels of derivation. It's not that directly all these scripts haven't got out from the Brahmi, but this heritage of Brahmi has bound all these separate scripts into one philosophy of encoding, and that is probably the reason why this panel is being formed as one panel. Geographically speaking, these are the scripts being used in Central Asia, South Asia, and Southeast Asia. The languages that use this script are multiple in number.

So there is one point that I would like to highlight about new Brahmis, like in case of Arabic script, every writing system has their own set of issues, if I may say so, like Arabic has one big core page along with some supplementary planes. Many languages are [worded out] out of that. So their problems are more related to one code block and many languages and different languages use similar perceptible characters in different forms. In case of Chinese, there are issues related to script undergoing revision, and then there are a couple of more scripts, and then there are language out of that. In Latin, again, there are a couple of core pages and languages [worded out] of that.

In case of Brahmi, there are multiple scripts, as well as multiple languages. Within the playoff between these languages and scripts is there are some languages which are returning multiple scripts, and

there are some scripts which cater to multiple languages as well. So when we are looking at LGR from this side, we have to take these things into consideration as well. In this particular presentation, I'm not going into the details of complexities in more than the Brahmi-based languages because we have already covered that, and if somebody is more interested in that, there are already reports out and I just wanted to give this particular presentation more about Generation Panels. So we're not going into that in detail.

This is just a depiction of how these languages are derived from the metrical Brahmi script. We can see the registration has the later kind of Brahmi – the way it is being represented. I'm just pointing on the board that this particular sign is the sign which was then for the later for the letter ka. Okay, I just forget to mention here the domain information here was taken from Wikipedia, so that reference is missing over there.

But there were three major families that [worded out] out of Brahmi. Those were Gupta [inaudible]. These are the leaf nodes of actual modern scripts that are being used, and from Brahmi there was a branch of Gupta writing system and out of that [inaudible] major families. Out of [inaudible] came the majority of the modern scripts that are used in India, and [inaudible] is one of the scripts which has been derived from [inaudible]. The way this thing goes is all the leaf nodes that have been shown here, those are the modern languages and this is their lineage.

Next slide – okay, I have the pointer. Sorry. Okay. Why Brahmi? Again, despite the variations – if I can go back the previous slide – we can see the ka character as it has been represented in Brahmi. It is majorly

different in each of the scripts. In the way they have been recorded into the Unicode, there are a separate code of locks for them as well.

Why they are common? It is because they are all having [inaudible] approach of having this level. They follow a specific structure and syntax, and this syntax is supposed to be presented in the form of whole label evaluation roots. We want all the whole label evaluation roots for all these scripts to be of the same philosophy.

UNIDENTIFIED MALE: [inaudible]

AKSHAT JOSHI: Okay, so why these whole label evaluation rules are so important is because when they're adhered to, they bring in some form of uniformity in the way the variants are analyzed. So that in a way is a security consideration when it comes to using domain names in these languages.

In the second-level IDN ccTLDs, they also are taking these things into consideration. Here we are dealing with the root, so those issues become even more prominent, and that's why we need them to be uniform across the board.

Okay, so why are there new Bramhis? Why not just the Brahmis? Because we are only catering to the modern set of languages that have come up out of the Brahmi family, and this approach is also in consonants with the ones that are already [inaudible] in principle of the LGR procedure. [inaudible] possibility of including the historic script has

anyways been excluded by the MSR approach, so that would have anyways been ensured, but then I just wanted to fly here that this approach concerning consonants with the [inaudible] principle.

So we're talking about the India person. I just wanted to give a glimpse of what are on the languages and scripts and which language family they belong to. So I'm not going to details of each of them, but then you can see that there are many languages which are shading the language families and scripts and can see the interplay between the languages and their representative scripts.

So we have an example of [inaudible] language over here, which can be returning [inaudible] Arabic and [inaudible] so it's like the [inaudible] language, or say for example [inaudible] will overlap with the Arabic Task Force as well, but currently we are not concentrating on that aspect because Arabic Task Force will be taking care of all those issues.

Apart from that, there are cultures outside India who have adopted the Brahmi script family to represent their languages. As for Unicode, they have been classified, again, centrally. They have been separated as like Central Asian, South Asian, and Southeast Asian. Some of them are Sinhala, which is the official language in Sri Lanka. I think it's script, I guess. I'm not really sure about the language as well.

Then there are Tibetan [inaudible], but currently the way MSR proposal goes, these four options are not part of the current LGR, but then we just wanted to flag that these are some of the similar languages. But it seems the Integration Panel has taken the view that they can be separated, so there are no issues –

UNIDENTIFED MALE: [inaudible]

AKSHAT JOSHI: Sinhala is part of the permissible set but Tibetan [inaudible] in Tibetan. [inaudible] are considered to be separate.

UNIDENTIFED MALE: [inaudible]

AKSHAT JOSHI: Okay.

DONG-MAN LEE: Yeah, we're just saying some of those we receive we have a mix here because some of them [inaudible], some [inaudible] Tibetan is not [inaudible] – what did we say?

AKSHAT JOSHI: Written in Tibetan.

DONG-MAN LEE: Tibetan. So it's [inaudible] Thai is in. [inaudible] is in. But it doesn't matter. They'll also be done by different LGRs anyways, so I don't think that's a concern.

AKSHAT JOSHI: Okay. So I'm going at first regarding these languages, we have identified some of the experts and we are reaching out to them, asking them to participate in the process.

One important thing that we are doing when interacting with them is familiarizing them with the LGR procedure. There are some challenges that we are facing with that, but I have a separate slide for that, so we'll see that.

What we currently seek is, apart from gathering a community within India, we also seek global participation because there are some of the other countries as well which are having some of the languages that are also used in India as official languages. It's not that we are only concentrating on the official language aspect of it. If there are users who are using languages which are not official but they are from different countries, we definitely want them on board because we'll get that perspective as well in our Generation Panels.

Some of them which we have identified in Singapore, Tamil is one of the official languages, so we will definitely be interested in getting feedback from Tamil community or participation rather than feedback from Tamil community in Singapore as well as Sri Lanka. There is a Nepali being one of the official languages in Nepal, Bengali in Bangladesh, and Hindi in Fiji. This is what we have identified. If there are more, we're definitely welcoming them.

UNIDENTIFIED FEMALE: [inaudible]

AKSHAT JOSHI: Sure.

UNIDENTIFIED FEMALE: Where does Rohingya lie? Rohingya is the language of a particular community of Myanmar or Burma, and also it's related to a community in Bangladesh if I'm not mistaken in the [inaudible] area. I just wanted to know whether –

AKSHAT JOSHI: Which language you're asking about?

UNIDENTIFIED FEMALE: Rohingya. Maybe the language experts can help us.

AKSHAT JOSHI: I'm not getting the language.

UNIDENTIFIED MALE: [inaudible] I'm going to assume it's more likely to be closer to [inaudible] than Arabic [inaudible] community in Myanmar.

AKSHAT JOSHI: Oh, you're asking what Myanmar? I guess it should be mostly be used in Myanmar. No, there is a specific script for Myanmar as well.

SARMAD HUSSAIN: [inaudible] the Muslim community that [inaudible]. Yes, there's a community that is Muslim community on the west bank of Myanmar, which is in fact has some political issues, pretty serious issues, without going to much in details.

It's my understanding in fact they do use Arabic script. That would be covered. In fact, another thing point when I was within the Arabic GP proposal, I did notice that if it was not covered or there was some points. It was not a major point, but that's basically [inaudible] covered by using the Arabic script.

AKSHAT JOSHI: Is this a good point to jump in or do you want me to finish?

UNIDENTIFIED FEMALE: No, no, it's okay.

SARMAD HUSSAIN: Okay. Just to follow up, I think, I guess a point to be made is that there may be many other languages which are using these scripts beyond. This is something you already suggested, but I think it will still be useful to explicitly reach out to many of those languages which are not "official," but are used by communities, and they may be using some of these code points in very different ways. That I guess perspective is good to have on board as well.

AKSHAT JOSHI: In the slide that we had there are actually 11 to 12 different languages that use the [inaudible] as a script. So those are anyways to be taken on board when we are finalizing the Generation Panel of the LGR person. Pardon?

NAELA SARRAS: There are two comments for Jacques in the chat room that went along with the discussions, so I apologize: Dhzongkha is the official language of Bhutan and shares a script based with Tibetan with certain differences. The other comment is: Myanmar is the official script of [inaudible] Myanmar, and [inaudible] recognizes it as such.

AKSHAT JOSHI: Okay. Going into my last slide, I guess, yes. Firstly, we've definitely been talking to Dr. Sarmad Hussain for his experience, so we are definitely learning from the way he has approached to the procedure of forming a task force on this and for the root LGR as well. So we hope to learn from him on how to approach this thing. We have been coming across some of the challenges when we are seeking people to be volunteering to this process.

One of them is the need for easy-to-explain formal documentation for the procedure and its expected outcome. What happens is we are already in this process, but now when we are reaching our bit of community, there are ten different languages. We share the [inaudible] script and now to get their perspective also we need to take them

onboard. We technically understand the script, but to get their perspective is also importance once we finalize something at this level.

So it is becoming mandatory that we explain then what we are trying to do, and since procedure is a bit complicated and this, at times, becomes difficult when we try to make them understand, and if they get away from that this is something that is complicated, acting as a volunteer for something like that at times is a deterrence. So even as an inter-panel communication, if any other panels have approaches like when they reached out the community, they had some form of explaining to the community that this is the procedure and this is how we do it. If there is any form of multimedia or easy-to-explain documentation that has been made, we will be also making that. We will also be sharing it with other panels, and if ICANN can also look into it in simplifying the procedure to be made understood by the people, that would also be very helpful.

Last point is, at times, we got a lack of funding related to this activity. We do understand it's purely community-based volunteer-based activity, but those people, even those volunteers who we are seeking or are there, they are in some way experts in those languages, and I don't know. There is definite support from ICANN related to the facilitation of the communication but this also is one of desperate is that comes to mind when somebody's asked to volunteer. With that, I would conclude my slides. Thank you.

NAELA SARRAS: Okay, thank you, Akshat, and again, tremendous effort going there. I wanted to tell you that I did hear I think several points that came

through in your presentation and perhaps very well-summarized, and I wanted to tell you that we have full support. This has come repeatedly in presentations in private meetings this week. We have full support from our management to try and facilitate this work to happen, so whether it's through requests for funding or providing better materials, doing more outreach, more like a more targeted, well-thought-of outreach plan with material to provide.

So I can promise you in the next few months, you should be able to see more from staff in terms of more outreach and more targeted stuff, really, and I want to assure you that we have the full support of forward-looking into funding issues. I'm not promising and I cannot promise anything, but I did want to echo several of our executives that have repeatedly said this is an important effort for ICANN and we have their support. So we'll see what we can do with that.

AKSHAT JOSHI: Sure. Any support would be definitely welcomed. Thank you.

NAELA SARRAS: Thank you. Go ahead, Marc.

MARC BLANCHET: Question and a comment. I'll start with the comment, which is essentially similar to what Naela says, which is that we've been hearing that ICANN really wants to support this work, so I would suggest as a personal comment – not IP or anything – please ask. Please ask ICANN

for support. Don't be shy. So Fadi Chehadé said clearly, so take his work and push what you need. At this point, only personal [inaudible].

Question is: My perception on the fact that you're taking the Brahmi scope makes things much larger, right? The languages, regions, countries, and so on; so it seems to me it becomes a larger challenge.

So on the funding side, how would you plan to address that large coverage in terms of giving all people, and how could you manage this task? Do you already have some ideas and ways you will manage this?

AKSHAT JOSHI: Actually, we have not yet formed a panel, so we do not have very clear-cut parts about what kinds of funding support would be needed. So maybe once we get to it, we'll have a more clear idea about that count.

The funding could be like if at all we plan to have a meeting among even within the Generation Panel members. I'm not really clear about those things, so I do not speak specifically on that. Maybe once we come to it, we'll have some more specific questions and we'll put up to ICANN sometime today.

NAELA SARRAS: Okay, that sounds fair. Okay, let's go that way. All right, because that was really informative, especially that one chart, that was really fascinating, so any other comments or questions for Akshat? I know we are due for a break – a deserved one – but I want to give Akshat his due time. So if we don't have any other comments or questions, what I propose we do is we take the break that we planned for this time. Let's

go for our full 15-minute break. It's about 3:04, 3:05 my time. Let's reconvene at 3:20, yes?

UNIDENTIFIED MALE: Maybe, Naela, you could state what are the agenda items for after the break so people will know what will happen.

NAELA SARRAS: Oh, yes. Thank you. Yes, because this was a really good session. So in the rest of the afternoon we're really going to switch gears a little bit, so we are going to present first more on the Maximal Starting Repertoire – a lot more detail than we discussed this morning. Then the last bit will be on the XML specification. I understand there is more interest in the XML specification, so we'll try to cover comprehensively but quickly the MSR and go into XML. All right, see you here at 3:20

Okay, if people could please come in and take their seats, we would like to get started. Okay. I'm going to go ahead and ask Asmus to take us to the next session designed today, which is to go over the Maximal Starting Repertoire. For everybody's benefit, we have a hard deadline at 5:00, so at least some of us want to go join the ALAC session. It will be here. That's even better.

UNIDENTIFIED MALE: [inaudible]

NAELA SARRAS: Thank you. So we're going to take from now until 3:50 to speak about the Maximal Starting Repertoire and the whole label evaluation rules, and we really want to protect the time for the XML because that is done at request of community members and we want to make sure we give that the time we promised for it.

So go ahead, Asmus, please.

ASMUS FREYTAG: Thank you. Given that the previous presentation ran a little over and given that the available time for the presentation was already very short, a few slides have to go through a very quick fast-forward mode so that we can recover at least some part of it.

The thing we were going to discuss here is the context that we have for development of the Maximal Starting Repertoire, and what we put in it. Since we're not many people here, I'd like a quick show of hands. Who was here in the morning session? Who was not here in the morning session? Two. Okay, I will very quickly go through the slides that are equal to the ones in the morning session and spend more time on the ones that are specific to the ones we want to talk here. That's I think the best use we can make of the short time.

This first slides are talking about the context of the Maximal Starting Repertoire, and the first important step to spend some time on this: what were we required to be put in as Integration Panel into the Maximal Starting Repertoire? The MSR is something that is supposed to be the sandbox for the Generation Panels to do their work, and so it defines the other limit.

The first outer limit for the enterprise is given by the fact that we are within the IDNA2008. That is the first cut that we are mandated to make from the Unicode repertoire. And the next cut is that the root does not contain any punctuation characters or digits, so we took out all punctuation characters and digits, otherwise IDNA2008 PVALID code points.

Then there were some specific prescriptions in the procedure to eliminate the characters requiring context rules under IDNA2008, so they went. And there are some scripts, and in some scripts individual characters for which the encoding in Unicode is not recently been stable. Those have been either permanently or temporarily excluded from the MSR.

Then finally, the procedure talks very clearly that the goal is to not support the scripts or code points for historic and obsolete users, nor for certain kinds of limited uses, which I'll go into some more detail later. Those we were forced to take out.

Any code point that ended up being outside the MSR cannot be part of the root zone label generation rule set. The MSR is, however, not the final LGR, so it contains some code points that may need further review, and can be possibly admitted if the review turns up a good justification for including that code point, or may be excluded if the review by the Generation Panel points to the fact that the code point is not acceptable after all.

There is the possibility that some code points may only be acceptable in particular situations. That is usually not a preferred outcome, with some

exceptions because it makes matters more complex, but there are some justifications that we may talk about where that might happen.

We talked about this: where we are with MSR-1 in the morning; that it's available for comment. For those weren't there, we do want comments by the 21st of May, and the public comments would result in us revising it potentially, depending on what kind of comments we're getting. But once it's finished, it is frozen. The sandbox gets locked in place, and the play can start. For those code points and scripts that are in principle eligible for the root but for various reasons were not included in the first version, there will be MSR-2 in reasonable short order to add those and make those available in a large sandbox.

The next step of the general procedure is that the Generation Panels go and take our MSR and go do the research and pick explicitly the subset that we need and justify their choices, but also add all the other pieces, like variants, whole label evaluation rules and whatever that are not part of the MSR.

As we've discussed here at length, the process should be driven in a way that it involves a good interaction between the panels, a good collaboration between the Integration Panel and the Generation Panels so that the Integration Panel receives things about which it has prior knowledge, is not surprised by it, and has the best possible chance to do a good job of reviewing and integrating the LGR proposals.

Here's a diagram. It's a little bit different from what we had this morning. How the information flow is doing the proposal of a script-based LGR. On the top left, we start with what the Generation Panel has

currently created, which is the Maximal Starting Repertoire, which have a line item there for these default whole label evaluation rules. So we have a list of code points, and we have a list of default rules how to exercise them and label validation.

That input is generic. It covers the entire root. That is taken by a Generation Panel. That's kind of the gray box, and it produces its work items are to finalize the permissible code points or repertoire for the specific script, which is a subset of the MSR which is first cut by, of course, the scope of the Generation Panel – the script that it has the scope for – and then further cut by removing any characters which have problems which are not really part of what should be permissible for the root. So the Integration Panel usually would expect that what the Generation Panel returns is a subset of what was present for a given script in the MSR.

If a script contains variants, then the Generation Panel would define the variant mappings, and it would for each mapping define a kind of disposition value, and it would give whole label evaluation rules if necessary to complete the set into a full-fledged LGR and those four elements define a label generation rule set. The idea is that if you have any label coming in in the machine processing of this label generation rule set, then at the end you can tell is the label valid? Is it not valid? You can tell does it have variants? Does it have not variants? If it has variants, are any of them allocatable? If not, the reminder of them would be blocked variants.

So as long as the LGR produced by the Generation Panel follows these rules, it can be submitted formally and the first step it goes into public

comment. When it's finished, it gets handed to the Integration Panel, which is shown in yellow again, and it evaluates whether the LGR can be integrated and if it is acceptable for integration, then it can be added to the integration root zone LGR, and if it's not acceptable for integration because of issues or because it can't get consensus in the Integration Panel, then it would get back to the Generation Panel for further evaluation, further work, and the process can take over. At the end, the Integration Panel will put to public comment the entire set of Integrated LGRs, and once they pass public comment, then they can be adopted for the root.

I'm going to be quick on this. We have shown this detail in the morning: what is in it in terms of numbers. What we haven't talked as much about is what type of code points are the ones that we are looking for in the root, and the best way to maybe describe it is what we're looking for is code points that can be used to cover modern use in everyday writing.

By "modern," we mean it has to be a currently living community that uses these code points. "Every day writing" means we're looking for things that businesses conducted in that writing system. It's not just a ceremonial writing system or one that is reserved for poetry or some other such uses. It has to be really every day, because there are some language communities in the world that use a language and use a writing system for certain parts of their cultural heritage or identity but actually in daily life use some other language and/or some other writing system to conduct their business. So we're looking for what is really in modern use for everyday writing.

In some cases, we have this issue of having these really small communities that use languages just like [inaudible]. You have many languages in the former Soviet Union that can have really, really small numbers of speakers, or Latin, we have small populations in Africa or the Pacific area. But we really don't think that mere population size is a good criterion because we have very active populations that can be very small. I'm thinking for instance of Iceland as a nation, which has about 300,000. They use some code points that are fairly specific just to their language, and it's a highly developed country and they're very vigorous and do all their business in that extension of the Latin alphabet that they're using. So it is not the population size that is the key factor. It is what you would call the effective demand for a writing system.

How can we do this? We researched this as part of the MSR because we wanted to construct the MSR so that it would cater to this kind of effective demand, and we came across the document called the Ethnologue. It's available online from SIL, which classified essentially all existing languages by both population and by type of use. The way they do the type of use, you can be characterized by the term "established vitality."

Now what the Ethnologue covers doesn't cover actually writing systems. It covers languages. But what we're thinking is if we know a particular code point is used by a language, then the established vitality of that language can be reasonable initial proxy, at least for purposes of the MSR, for suspected effective demand for the associated writing system. It is not always a one-to-one. It's some writing systems that are not in strong demand because the languages may be vigorous, but they don't write. So it's only a proxy.

For the MSR, the Integration Panel looked at the classification from the EGIDS (the Expanded Graded Intergenerational Disruption Scale), which covers things like whether language is about to die out because they have no children speaking it and stuff like that. So we looked at the different levels they defined for that, and there was level four and level five, and in between those two, with the higher numbers being the less in-demand languages, we made a general cutoff for the purpose of adding code points to the MSR.

MATT STOFBERG: But you were saying that four is enough.

ASMUS FREYTAG: Would you speak into your microphone and identify yourself?

MATT STOFBERG: Matt Stofberg, .sa. Do you say that level four is enough to be included, or is level five required?

ASMSUS FREYTAG: Sorry. What I just said is that the Integration Panel made the cutoff between levels four and five. So in our opinion, we think that a language that is in vigorous use today which has a [inaudible] of standard as formed but it is neither widespread or sustainable, it's not a language that is on the must-support level of the root unless we hear otherwise with good arguments and public comment.

MATT STOFBERG: Okay.

ASMSUS FREYTAG We put our best face forward in what we put together, and it's all subject to public comment. And if you look at the particular language and say, no, we got it wrong, then tell us so. Then we can do something about it. Another question?

UNIDENTIFIED MALE: If some of these languages rejuvenated at some point, it's by taking them out of MSR, it means that they can never come back in. So isn't it better to keep them in MSR but not allow them in LGR?

ASMUS FREYTAG: That is making an assumption that we haven't said anywhere. It is not stated anywhere that MSR-2 cannot be a superset of MSR-1. I'm going to try to keep this short. I have almost no time to finish my remaining slides. I think it's not going to be the issue you think because you made the assumption that we're going to not be able to create differently-sized sandboxes in a few years, when languages have changed, when new code points have been added. That can all be adjusted to.

In creating the MSR, the Integration Panel started with a list of scripts and then looked at, for those scripts that are in vigorous modern use, the reasons why certain code points should be excluded and for instance, where they're purely historic, limited specialized use. There's

many code points in Unicode that exist solely to do phonetic notation, which is not necessary for [inaudible] labels, and there are, however, code points for which it's not clear whether they're historic only. It's not clear whether they're not really used in everyday writing, even though they're also used for phonetics. Where that uncertainty was known, we as Integration Panel decided to leave these code points in the MSR because it is up to the Generation Panels to do the detailed study on them. If we overshot our target, then it's for the public comment to help us add certain code points back into the pool of those that need to be reviewed by the Generation Panel.

NAELA SARRAS: Asmus, I need to do a time check. So it's 3:43 and I know you want to be conscientious about the time, so I feel like you want to flesh this out a little bit more still.

ASMUS FREYTAG: I would like to ask that we can take questions at the end because otherwise we will have to stop in the middle of some random slide and not go to the –

NAELA SARRAS: What I'm going to do is I'm going to give you exactly until 3:50 to finish the slides, which is not a small task, and then we're going to hear from Sarmad and the online –

ASMUS FREYTAG: Good. I think that's a good way to do it. So what is listed here is what is also listed on the documents that accompany MSR, so I'm not going to walk through all of these. It's basically a classification of the top reasons why certain code points were considered for exclusion from the MSR, and in fact in the annotated co-table document that we have as part of the MSR, you can see which character was excluded for what reason.

Here is the way this would look. For instance, here you have a piece of the annotated co-table for Greek. You have the Greek small letter zeta, which is given a code position of 0371, and we have annotated it as obsolete, so we have the information that is actually not used in current, modern everyday writing. 0375, if you look a little bit below at the lower numerical sign has a context O of rule in IDNA2008. It is also obsolete, and for both of those reasons, it has been excluded from the MSR.

So as you go through the tables for all the other scripts, you find similar annotations where the way it works is if something is highlighted in pink, it is IDNA2008 PVALID, but excluded from the MSR. If it is white, it is not PVALID or it is a digit, and if it is in yellow, it is what is in the current public comment version of the MSR.

One thing that we did not discuss in the morning but is really important has to do with combining sequences. There are many scripts that use combining characters that require a base character to combine into another form. In some scripts like Latin, you have pre-composed forms and IDNA2008 requiring normalization form, and if C means that when the pre-composed form exists, you have to use it. But there are a number of languages use marks for which no pre-composed forms exist.

There are some scripts that use combining marks with no pre-composed forms, and there are differences in those sets. Some of them are optional in some scripts and languages and some are less or not optional.

There's one block in the Unicode stand that the combining marks, the general ones, those are mainly used in the context of Latin, Greek, and Cyrillic. In Greek, there's only three that are being used, and all the combinations that exist for them are pre-composed and would be required to be used pre-composed, so we don't need to add any combining marks themselves for the Greek repertoire.

In Latin and Cyrillic, there are some marks such as vertical bar below that is used in the [inaudible] languages for which there is no pre-composed forms, and in that case, the Integration Panel, and in similar cases, has added the combining mark to the MSR. It has also added the combining mark if it was used in a combination on the idea that those marks that show up in combinations are frequently used, and there may be unrecorded combinations that they are being used for in some languages, and we wanted the Generation Panels to be able to investigate those combinations, so we left those combing marks in.

So we expect the Generation Panels to do this investigation and in some cases where there's only a limited number of fixed combinations that are allowed to not simply allow the combining mark into the repertoire, but maybe only allow it as part of a fixed combination.

In various other scripts, the rules for combining marks are very different, and there are several kinds of options a Generation Panel may

consider it. If marks have an optional character, they might not be supported all together. If they occur rarely, they might be supported in fixed combinations. If they occur as part of well-formed clusters, then maybe there might be a whole label evaluation rule that defines what a [inaudible] cluster is. There are cases where they will just have to be treated like ordinary code points and just be part of the repertoire.

Another part of the repertoire, which gave us particular challenge, was the Han ideographs because they have no easy division into common, modern every day and historic, and so the Integration Panel started with a superset of existing IDN tables and made a superset of that and the IR core repertoire and created from that an outer limit that is part of the draft MSR for review.

The Han tables are similarly printed and annotated here. Some example: one it shows a number of characters. The yellow ones would be the ones that are supported in the MSR. For each of them, it's given an indication of which of the groups – Chinese, Japanese, Korean – that Han character is used in.

The Generation Panels have to evaluate whether to add variants. We discussed that bit in the morning. And they have to discuss whether to put whole label evaluation rules into the LGR. Whole label evaluation rules can be used to filter out in some complex scripts. If you put a random sequence of code point together, you can create something that can maliciously break displays, and that would be the intended use that whole label evaluation rules were put in for into the procedure to allow such stuff to be filtered, but it could also be potentially used to

fine tune the set of allowable allocatable variants to reduce that and minimize that.

The MSR-1 contains a single default WLE rule, which is intended to prevent labels with leading combining marks.

We're going to have a whole tutorial on the XML format. Just to make the connection this and the MSR and the LGR work is that the MSR itself has been delivered. The normative specification of it is an XML file, according to this format. The format itself is not specific to the roots on LGR, so it has many features that will not be used in the root zone work, and the Integration Panel is preparing a document that shows which features of the general format will be part of the work for the root zone.

But if somebody has a very simple script, the typical situation might be they can just go take the MSR, delete all the repertoire lines that are not part for their script, and update the language stamp and update the description, and they'll have an XML file without having to do a lot of editing.

But for any more complex cases, the Integration Panel or ICANN will certainly want to assist and help Generation Panels with the proctor aspects of preparing their submission that uses this format.

NAELA SARRAS: Thank you, Asmus. So we go to you, Sarmad, and then we will take questions in the online group.

SARMAD HUSSAIN: This is just following up Asmus' comment, and I think if that is really the case, I think it's very useful to document that because that has a lot of implications on our current work.

I think what you were suggesting was – and again, at a personal level, that was not my understanding at this time – and that was that MSR-2 actually includes some of the characters that are pink in MSR-1.

ASMSUS FREYTAG: Yes, so let's be clear. We know languages change over time. That's just a reality. We know written form of languages change, and in those situations where the languages are evolving – for instance they are becoming more widely adopted or a writing system has been adopted by a language community – and it requires a code point that was previously, say, only used for historic purpose, there's absolutely no reason why future version of the MSR and a future version of the LGR can't in principle add support in an additive manner for those characters or code points that were previously limited to historical use.

I say "in principle" because the one thing we will have to do is to make sure that whenever we allow such addition that we can do that in a compatible way. But assuming compatibility is not an issue, there is absolutely no problem to react to new information or new developments.

NAELA SARRAS: Are there any questions?

UNIDENTIFIED MALE: We have one question in the Adobe room: The cutoff between Ethnologue levels four and five implies that you either don't recognize or are deliberately rejecting the potential the DNS may have for offsetting the forces of language attrition. To what extent was that discussed before setting this limitation?

ASMUS FREYTAG: We had a very careful discussion of individual languages, and we were following up on the available information for many of the languages beyond just simply relying on the Ethnologue. It is certainly not the intent of the Integration Panel to blindly force cutoffs that result in languages using useful support, but if you look at the individual languages that we evaluated for these cutoffs, in those cases that they use unique code points – it only matters if the language uses a unique code point that's not used by any other language already – we're not really aware of any particular case where that cutoff results in denying access to a really vibrant and active language community.

NAELA SARRAS: Is that it from online? Yes?

UNIDENTIFIED MALE: Yes.

NAELA SARRAS: Yes. Thank you. Okay, are there any other questions in the room? Comments? Behind me. Okay, so I think we can safely switch to the next

topic. We're calling it "Training on XML Format to Represent the LGR." Let's go ahead and hear from Kim Davies.

KIM DAVIES: Thanks. For a start, I just wanted to acknowledge the contributions of Asmus and Wil recently in developing this standard. Particularly, Asmus has added has added a lot of the whole label evaluation material to the specification, and Wil has helped us a lot with evolving the schema and doing testing and so on. So it's definitely been a joint effort in developing this.

Pretty much what it says on the screen: we'll talk about why we created this format, what the main sort of processing logical pieces are of the format and the various processing steps that can be performed.

Much like Asmus earlier, I think almost everybody in this room heard the distilled version of this morning, so I probably don't need to belabor the basics too much, but in essence, this is about creating a format that could accurately represent label generation rule sets. If we go back in history, we used to call them IDN tables. In recognition of the fact that they're more than simple tables and they support this root LGR, we've started using the terminology "label generation rule set."

Primarily, these rules are used for determining label eligibility, but they can also be used for determining variants. In terms of designing the format, we had multiple choices. The existing IDN tables that are out there are essentially CSV files or something comparable to that.

We settled on XML for a few reasons. Firstly, it's a format that has a lot of tool chains out there. You can do things like validation. You have editors that are specialized in editing an XML. It's machine-readable, which is essential for us to fulfill our ambition that this format can be read by software implemented in a consistent way. It is a format that we believe can represent all the existing IDN tables that are out there. It's a format that enables us to explicitly identify the policy rules that apply to label generation.

As you already know, this is the format that we settled upon for the root LGR, so whilst the application of this particular format is beyond simply just the root LGR and a lot of the complexity in the format isn't necessarily directly relevant to the work on the root LGR, it is the intent to use this for the root LGR and for other domains.

Sorry. In essence, at a high level, how does this work? There's two pieces to using the LGR. One is the file format – the table – the other is the tool. The tool takes in the LGR file. It takes in some kind of input such as a domain, and it spits out a result. The result is such as this domain is eligible for registration, or this domain is not eligible for registration.

If we look at a slightly more complex example, if you define variants, you can similarly put in input of a label and it would generate multiple outputs. You can also tag certain evaluations with an action. These actions help define registry policy. For example, certain combinations of the variants might be allocatable. Others might be blockable. Depending on the linguistic rules that have been codified in an LGR, you can define multiple outcomes by some of the various combinations.

Sorry. If we break down what an LGR is comprised of, it's comprised of firstly lists of code points. This is what you'd typically expect to see in an IDN table today. It also includes variant code point mappings, so if there's a direct correlation between a code point and some kind of variant code point, you can list that. That's what we see, for example, in the CJK table.

Additional things that the LGR can represent that existing IDN tables typically don't is firstly character classes, the ability to create groups of code points and treat them in certain ways. We also have the ability to derive those classes from things like Unicode properties. So rather than enumerating a list of code points, we can use the knowledge inside the Unicode character database to greatly simplify a lot of situations for LGRs.

We also have the capacity to do whole label evaluation rules. In essence, these are contextual rules, essentially saying certain permutations; certain code points are eligible in certain conditions, but not all. So rather than simply saying, "This code point is allowed anywhere in this string," we can say, "This code point is allowed when following this kind of character when it appears at the start of a string," and so on. So we can really add nuance to the rules and add a lot of awareness of linguistically how certain scripts are used. We know that a lot of the various elements of scripts only appear in certain context, and these whole label evaluation rules allow you to express that.

Finally, we have a section in the standard for actions. I mentioned before you can tag things like allocate block. Essentially, a lot of this was implicit in existing IDN tables that are out there, but you can explicitly

declare that, for example, if certain code points appear that are tagged in a certain way, that automatically means you block the string.

In the context of CJK, you could write a rule that says, “If all the code points in this string are tagged as simplified, you then put it in the zone. If all the code points in this string are tagged as traditional, you block, and if there’s a mixture of code points – some are traditional, some are simplified – that’s ineligible.” So you can do that kind of thing using action rules.

Just to go into a little bit more detail, code point lists is really as it says a list enumerating code points, and you can optionally connect to those code point variants. The code point sets, as we discussed, can either be lists or they can be derived from Unicode properties. The example on the screen is you can simply say, “This code point list is all of the Greek script characters in Unicode.” It’s just as simple as that. So in one line, you can create a group of characters based on the Greek property, rather than going through enumerating a list, spelling it out. I think for a lot of these cases, by using that, you can actually greatly simplify the complexity of an LGR. We talked about whole label rules and we talked about actions.

Let’s look at the details. I don’t know if I can assume familiarity with XML from most people in the room. Suffice it to say, XML is a text-based format. It has a certain structure to it. It’s similar to HTML. We had a prescribed format in the specification. In essence, you put together an LGR following the specification. It starts off with an LGR tag. It then has three major sections. Firstly there’s a meta section. This section is essentially the meta data. None of it is particularly normative except

one element. There's a header for Unicode versions, so if there's a specific Unicode version that's needed for your LGR, you can specify it in there, and a compliant implementation of the standard should check to make sure it has the Unicode data from that specific version to operate against.

But other than that, this is where you put the contact details of the author. You can write down what domain names it's used for and so on. So this is really just where you put all the auxiliary data to help you interpret the table. The second section is the data section. This is where you enumerate all the code points. You have those lists. Then finally we have the rules section. This is where you define things like whole label evaluation, actions, and so on. Probably don't need to belabor the meta section too hard. It should be all relatively straightforward. In terms of the language tag, we use existing standards there, so you tag based on language or scripts based on those existing ISO standards. I think it's relatively self-explanatory.

In terms of defining code points, we use these [inaudible] tag to list characters. We use essentially Unicode notation for the code points without the U+ prefixing the hexadecimal. You can specify ranges. So if you for example allow the letters A through Z in Latin. You don't have to spell out 26 code points. You can just A to Z literally.

You could also specify sequences. For example, in [inaudible] the middle dot is already permissible when it's proceeded by an L and following an L, so you don't have to permit a middle dot in any context. You can just simply say, "L middle dot L is a permissible sequence," and that would be interpreted and permitted.

Just a simplified working example, here's a declaration of a particular code point. It has two variants, one of which has a disposition of allocate. The other has a disposition of block. So in practice, this reflects sort of a typical Chinese example where you have a traditional and simplified equivalence.

Following the XML snippet at the top, if you put this code point into an LGR tool, you would result with an output of two possibilities. One of those code points would be allocate, and the other one would be block. It's relatively simple.

Here's a slightly more fleshed out example. You can have more intricate interdependencies. You can specify that if the input string has certain characters that it has certain activities based on different rules. The point here is sort of these relationships don't have to necessarily be symmetric. They might depend on the input string.

ASMUS FREYTAG: [inaudible]

KIM DAVIES: Sorry, yes?

ASMUS FREYTAG: This is a beautiful slide to cut in and note something that has come up in discussions all week, and maybe it's useful to highlight. Notice that the way in this example diagram that they status information for the variant relationships is always placed on the error. That is, it is not placed on

the target of the variant relationship, but on the mapping itself. So the directionality for instance on the right, it matters whether you go from one character to the other or back. You get different status values. This is a really important part of the scheme that you go into variant relationships you have to consider. I just wanted to point it out because it's so nicely visible in this particular slide. It leads in ultimate to cases where it make perfect sense to map a code point to itself in the bottom left, just so that you have an error that you can give a disposition to under the scheme.

UNIDENTIFIED MALE: [inaudible] an explanation of why this is not an example of a consistent set of rules if it is because we've got the situation where the top one is blocking the one at the bottom on the left, but through an indirect chain, it validates the one at the bottom on the right, and that one will then validate that same one, which is already blocked, so how does that work? Or is it [inaudible]

UNIDENTIFED MALE: Do you have the answer, Kim?

UNIDENTIFIED MALE: You might have a better answer.

UNIDENTIFED MALE: I was going to try.

UNIDENTIFIED MALE: This is not meant to be a real example, it's really just to illustrate that it can go that way.

ASMUS FREYTAG: The first one to understand is that these things don't chain, okay? In the diagram, it looks like you can hop from one to another. When you evaluate a label for possible variant labels, you make only one hop at a time. This is very important. You don't get to navigate across the entire diagram. So you might have that blue thing at the top as original code point in the label, and when you create the permutation of variants, you can follow the green arrow down to the green one in one iteration, and the next iteration you can follow the red arrow down to the magenta one, and then you're done for that original label. Okay?

I don't know, Kim, if you have a better one.

KIM DAVIES: I was trying to think of a comparable example that's in something I could read.

UNIDENTIFED MALE: [inaudible] that's the problem.

KIM DAVIES: Right. Some of this complexity is only necessary for certain applications. But for example, if you had café with an accent, you might want to map

that with café without an accent, but you wouldn't want it to go in the opposite direction. Similarly, if you had a sharp S in German, you might want to map that to SS in Latin, but you definitely wouldn't want SS to always map to an S set.

So there are certain situations where you would want to have that conversion happen in one direction but not necessarily in the other direction.

UNIDENTIFIED MALE: Certainly not allocatable [inaudible]

KIM DAVIES: The mapping always happens in all directions. So the mappings exists but what you do based on them can change.

UNIDENTIFED MALE: Right.

MATT STOFBERG: I have another question. Matt Stofberg, .se. Should I interpret the rules there that 62E is never allocated, is never used? So that you have to point that the code point itself allocated? Or what is the interpretation of the rules?

ASMUS FREYTAG: Don't we have that on the next slide, Kim?

KIM DAVIES: Yeah, I think we do.

MATT STOFBERG: Because certainly the third example then it's allocating itself.

ASMUS FREYTAG: So here you have a label coming in and then Kim can run us through.

KIM DAVIES: Yes, so with this example, if your input string is the blue, the green, the purple at the top, and you have this set of rules, yeah, if you provide three of those blue characters, that would result in block. If you have two blues, one green, if you follow the arrows, it results in a block. In fact, the only combination that would result in allocation would be the green, the purple, the purple.

ASMUS FREYTAG: Actually the bottom ones are not the new labels, so they're all the possible permutations of the variants. Right? That's how we contracted our diagram. So the top is what comes in as the allocation. The bottom is all the possible permutation of variants that you can create, and the only one of those variants turns out to be allocatable.

KIM DAVIES: And not the input.

ASMUS FREYTAG: And the possible may well be that you always allow the input to be allocatable.

KIM DAVIES: Right. So in the actions declaration, you spell out those kinds of parameters. It's a matter of registry policy.

So in terms of the rules section, in essence you define things like the character classes that I mentioned earlier. You mention those whole label rules and actions. These allow you to define things like which labels are valid, well-formedness for whatever you might define that to be, and we just stepped through an example that is probably very RFC3743 like, but you can do those kind of things as well.

The default actions is basically the common case that if you match a variant, it's allocatable, and if not, it's not allowed. But you can be more explicit if you wish.

In essence, there's a lot of flexibility in the standard. You can do a lot of things. In fact, one of our design criteria for it is that we could actually express all of the IDNA2008 contextual rules as LGR. I think we've succeeded in that. So in essence, to write an implementation of the specification, it doesn't actually need a lot of special cases or [inaudible] to bake in certain criteria. It can have a very naïve approach to IDNs, and you basically spell out the roles inside the LGR.

So whole label evaluation roles: this is probably the most potentially complex part of the specification. If you're familiar with regular

expressions, it's a bit like that. Regular expressions are rules for matching characters. It's not a regular expression-style syntax. Just conceptually it's like regular expressions. Obviously, we're using an XML format here.

So for example, the rule here is to – it's given the name Leading on Spacing Mark [?], so perhaps it's obvious what it's trying to catch – the rule essentially says the start tag means apply this at the start of a string, and it's followed by matching character of the class. GC is MN, which happens to match. General class I think is a non-spacing mark.

This is an illustration of actually leveraging the Unicode specification. If you didn't do it this way, you actually have to go into the Unicode specification, find all the code point values that match as non-spacing marks, and list them out manually. Here you can actually leverage the Unicode specification to create a relatively simple role.

We mentioned earlier you could also create classes of code points. So here you can make a class of digits. The numbers 0 through nine matched to 30 through 39 in the Unicode spec. Create a Unicode properties as we just mentioned.

You can also tag code points. For example, I mentioned traditional and simplified before. In the complete list of a CJK table, you can add tags to individual code points and explicitly declare this is a simplified code point. This is traditional. And then later on, you can use those tags in your rules. You can say earlier [inaudible] code points of a simplified tag and don't allow it to comingle with the traditional tag, for example.

You can also set operations. For those familiar with that, it means you can have multiple sets and combine them. You can negate them. You can perform all the kind of standard set operations.

In terms of defining these regular expression [inaudible] constructions, there's a start tag which matches to the start of a string. I guess that would be like a [inaudible] in a regular expression. An end tag, which would match to a dollar sign in a regular expression. Any, which would match to a dot in a regular expression. We can match on [inaudible] on code points. We can match to a class. We can match to a choice. So if you wanted to match on either one code point or another, or even one set or another, you can basically do a choice, which is either/or – an or operator.

You can do look behinds and look aheads. We saw in the IDNA contextual rules it only allotted certain code points if previous to them certain conditions were met, and after it, certain conditions were met. So you can use look behind and look ahead to represent that.

As noted at the bottom, essentially have analogs in regular expressions. As we've been developing the specification, we've been trying this out in [inaudible] ourselves. I know Wil has actually gone to, in the implementation he's been working on, that's literally what he does in his code. He reads the XML, maps it to a regular expression, and uses the underlying regular expression engine to do the actual testing.

Here's an example of using context in a rule. Here you can have character – this code point 200D. It's only valid when the condition is satisfied of the rule follows [inaudible]. Also in the document, the

second diagram, we defined what follows [inaudible] means. There's a look behind rule, so it's saying, "Prior to this code point, this rule needs to be satisfied. The way the rule is constructed, it needs to be Unicode property CCC is [inaudible]." I don't expect you to know what that means, but essentially that is the definition of a [inaudible] in the Unicode character database.

An implementation of this is basically: check if the character behind is a [inaudible] If so, it's satisfied.

Dispositions. These are actions. We mentioned earlier that you can define the consequences of calculating certain permutations. It doesn't always have to be allocate. You can list the set of rules. These rules basically say, "If it matches certain conditions, you can allocate. You can block. You can activate, or you can map it is invalid." They're the four verbs that are specified.

If in your operations you need to use more verbs, that possible. In the context of the root LGR, I think it's constrained. But it does provide additional flexibility. In registry operations, you did something else.

What triggers an action? You have rules, and if the rules are matched or not matched you can trigger an action. You can trigger actions based on the appearance of variants with certain dispositions. When a label exclusively contains variants or certain disposition or when all variants in a label are of a given disposition, so say that these are some of the possibilities of when you might trigger actions.

I think we have an illustration to show that. The first action is you can, if any of the variants are sort of tagged as blocked like any code point,

then automatically you can block the whole string. So it doesn't have to be all the code points are blocked. If any of the code points are ineligible, then you should block the string. So that's the first action.

The second rule is, if every single code point in a rule results in an allocate, then your result is to allocate. Then finally, the last rule is essentially saying any valid code point in this table is allocatable.

So these are some of the options on actions that you can implement. These are the default actions for a table, but you can specific alternate actions if you so wish.

In summary, label generation rule sets are really here to advance the way we represent IDN tables and provide you with new flexibility that isn't currently available in IDN tables. Whilst the specification is potentially unwieldy, and admittedly this was a very scattershot, two or three specification, I think to fully understand it, you probably want to read it in more depth.

We did expose a lot of complexity that you probably don't actually need. We're really trying to create a comprehensive specification that can do a lot of things, but in most normal cases, you don't actually need to use that complexity.

So don't be scared from the specification if a lot of this seemed confusing to you. At its heart, it's a way of structuring IDN table as you know them today in a machine-readable way that can be validated, and if all your table is is essentially a list of code points, you can represent that in a relatively simple way with only minor modifications to the way you compile IDN tables today.

You can add to it and add some slightly additional functionality in terms of contextual rules. Our sense is that whatever contextual rules are likely to be implemented are going to be only a small number of rules. It's not that you're going to be creating dozens or hundreds of rules. It's going to be very specific kind of contextual rules that in practice LGR is likely to contain.

But the key point is this is designed to be a universal format. Ultimately, we want all IDN tables, all LGRs to follow this format. The root LGR work is going to create a series of tables. It will be integrated into the root LGR.

But ultimately, we want to the whole ecosystem of IDN implementers our there to use this format. I mentioned this morning it provides us ultimately with a way to implement this in various different ways that should be easier for IDN implementers in the long run.

Just in terms of the resources available, firstly there is a schema for this XML. What a schema does is it formerly defines XML rules. One of the benefits of using XML is that it provides us with this faculty. It means that you can type out an LGR, and if you make typos, if you put certain tags in the wrong place, when you run it through schema validate, it will tell you. It's not going to catch everything, but it's going to catch a number of common errors that are likely to creep in.

The schema is published online. You can get to it right now. We publish it on GitHub. It's still evolving, but it's available now, so if you wanted to start authoring your own LGRs, you can pull the most recent version of the schema and use that.

We have additional resources on GitHub. This is a specification that's in draft. The specification is on GitHub. There's two documents that have been written – sort of supporting documents that I think are of interest. Firstly there is a document that explicitly says how you would convert an LGR into regular expressions. This is more for implementers of code – how they would do that.

But if you're familiar with regular expressions but you don't get what I just said, you can actually use it backwards. You can go in there and say, "I know in my head what a regular expression would look like." You can then see what the correlating declarations would be in XML.

We also explained how those IDNA2008 contextual rules would be represented in the LGR. So again, I mentioned that one of our tests that the LGR specification is sufficiently expressive was to look at all those contextual rules that already exist in IDNA and make sure that they could be represented in the LGR format. So this document spells out what those rules actually are. That's it. Very happy to take questions or drill down into things I might have glossed over too quickly.

NAELA SARRAS: Thanks, Kim. Let's go to Akshat.

AKSHAT JOSHI: Hi, Kim. First of all, it appears to be really a complement to a study of what we can see in the future LGR, and the work you have put into it really seems to be really good.

I just had a couple of comments, actually. One was regarding the general structure that you showed of the proposed XML wherein you showed allowed and disallowed and that field was mandatory. Can we go back to that? It was in one of the initial slides.

KIM DAVIES: It's taking a while for my clicks to register. Is this the slide?

AKSHAT JOSHI: Yes, yes, yes. So in this particular slide you're saying that it's mandatory – okay, that's just mandatory – but are they allowed or disallowed? So the way we have defined in the procedure, everything is by default excluded and inclusion is very specific. So instead of maintaining a list of disallowed code points, if we can modify to only contain those which are allowed, that would be a more neater approach, I guess.

KIM DAVIES: Yes, sorry. An empty table means nothing is allowed, so it's a model of inclusion. As you would list code points, they become allowed.

ASMUS FREYTAG: Yeah. Kim, I think this is simply a minor flaw in this slide wording. We meant to say allowed/disallowed as basically saying by listing it you allow it. But it can read as if there was a method to list something with the purpose of disallowing it in the code table, and that is in fact not possible. So by listing it you allow it and by not listing it you disallow it.

If we did this ten times in a row, we'd find all those.

AKSHAT JOSHI: Okay. One additional question was: I don't know in what form the LGRs would be made in. Is it like the Generation Panel would be giving this XML as such, or will we just giving the rules in the text form?

ASMSUS FREYTAG: Yeah, that may be more specific to the root zone, so I may field that question.

The Integration Panel does insist that the formal specification of the LGR be in the form of an XML file according to this format. That's very important so that there is not interpretation that we need to do on what precisely you mean. This format is unambiguous, and so we ask you to use it.

However, it is to be accompanied by a plain English document describing what you did and why you did it, and it's entirely appropriate to express the purpose of, say, whole label evaluation rule in English and running text as well so that we can be sure to understand why you choose the form. But the formal specification is the XML.

If you look at the Akshar example in the draft of the specification, you find that there is an example of where first there is a discussion of it in English, and then there's the XML format. When you have both, you can follow it. Similar for the Integration Panel, if you submit something, the rationale and overview document that is part of LGR submission should cover everything you put in XML format to help understand not only the

Integration Panel to help understand it, but also to help any public review of the LGR to understand what you put in there.

AKSHAT JOSHI: If I think it from the programmer's point of view, if I have some XML to deal with in my program, which would be built by somebody else, having this schema to validate that is okay, but once we do schema validation, I don't think we get a very specific answer of where the flaw lies.

If it is almost like whether it is valid or invalid, right?

KIM DAVIES: It depends on what kind of flaw you have in mind. Schema validation will check the syntax errors. It won't check that the logic of your rules makes sense. But my take on that is that as a designer of an LGR, you should have some test cases and corner cases in mind. Ultimately, we can then use these tools to make sure that all the things you think shouldn't be allowed to fail and all the things you think should be allowed to succeed, and you can run those test cases through your rules to make sure all your assumptions hold.

ASMUS FREYTAG: And not only that. The Integration Panel in its draft document called Requirements for Submission does say if you add whole label evaluation rules, we do expect that test cases are delivered in a separate file so we can make sure that your rule actually works the way you think it does. So if you develop test cases that you have verified, it should work with

your rule. You submit them with that. It has a secondary purpose, too, because some of the rules can be – well, in all programs, you can write equivalent statements using slightly different syntax elements or a slightly different order of syntax elements.

The same is true here. You can restate certain things using different syntax elements for precisely the same answer, and the Integration Panel would reserve the right to have the final version for public use possibly be a restated one that is functionality equivalent, and we can only do that if we have test cases to make sure that we don't mess up that process. I'm not saying we would, but we reserve the right to do that.

AKSHAT JOSHI: Actually, I was approaching it from a different angle, like here we are the internationalized character set, and one of our experiences when dealing with that is that there are many possible tools with which we can build this, like one can us Notepad or Notepad++ or a doc or anything like that. There are various inquiring schemes that are employed by these applications.

So many a times, when we get such files from different people, we get different inquiring schemes make the wrong character has been inserted in some and not the others. Those kinds of issues may crop up if we let the handling of the actual XML to the people.

Why I'm making this point is I have a suggestion in mind. If we can do that – make leaving the inquiring and the actual structuring of the XML to be done manually by the Generation Panel, can there be an interface

which only the generation has to fill in, and encoding to XML in the background?

ASMUS FREYTAG: No, we don't have a nice editing tool like that. I'm a little less worried about the encoding aspect because there's no reason why the XML files would have to use anything but ASCII characters because the descriptions are supposed to be in English. The detailed description should not be in the XML format. It should be in a separate word document and be free text. Your comments should also be limited to English, being the common language.

When it comes to further detailed support, we are certainly open to suggestions if there's anything that we can provide, short of having an editing interface. That sounds a little bit more resources than we can bring to that task.

But for instance, inside the Integration Panel we have separate tools that people have written for internal consumption, and we're using tools that we want to use for verification and merging, etc. I could see a possibility that somebody submits a draft version of a table as part of the dialogue between the Generation Panel and the Integration Panel. We run it through our tools and report back to you, saying, "Hey, it doesn't fly in our tools, so let's get together and see where the problem is."

But we're not interested in getting a large number of broken files when we have to do our integration work, so we're going to be smart about this and work with you.

KIM DAVIES: Just to add to that, the wide variety of file formats that are used to describe IDN tables is precisely one of the problems that result in us wanting to undertake this project in the first place.

I have a different hat to Asmus in that my primary motivator is the IDN repository and what we do at the second level of the DNS. Certainly in the context of the new gTLD program where all the applicants had to submit IDN tables and all the existing TLDs that have done so well there's a wide variety of formats, and they weren't consistent. So using XML and formally specifying the format was precisely designed to be a solution to that.

XML has certain properties that are beneficial. You talked about different character encodings. Well, XML is mandatory UTF-8, so it's simply not possible not to use UTF-8 in an XML file.

Notwithstanding the root LGR's ambition to be all in English, but theoretically, if you were producing an LGR for yourself, you can put UTF-8 descriptions in comment fields and what have you. That's perfectly fine. You could do that. The format would support that. But you couldn't save an XML file in anything other than UTF-8, and I would assume the validating tool would probably flag that such-and-such a string is not a valid UTF-8 string.

NAELA SARRAS: Asmus, I need to be able to go to the online participants and see if there is any other discussion here, yeah? So let's go to online please.

UNIDENTIFIED MALE: Yes. The question is: are LGR authors are expected to represent IDNA2008 validation, such as context zero or 0, or context J [inaudible] and hyphen rules in their XML files?

NAELA SARRAS: Who's asking the question?

UNIDENTIFIED MALE: Oh. James Mitchell from ARI Registry Services.

NAELA SARRAS: Yeah, go ahead Kim. Sorry. Asmus.

ASMSUS FREYTAG: It's certainly possible to create these rules in an XML format. It's been verified that that can be done. Also we've contacted some of the authors of the IDNA specification to have them verify that we have captured the intent according to their insight correctly. Whether in practice you include these rules in the XML files is something to be seen because it could be that we do something like develop a piece of XML that has these in canned form and then they get included into actual LGR files because they're common piece.

We haven't fully investigated the inclusion mechanisms, but we have been evolving the specification, taking pains to design it in a way that makes inclusion easier whenever there was a dependency on text

around the possibly-included block. We have tried to limit that so it's easier to include blocks, as the Lego bricks for common things.

For example, in the root zone work, the root zone will satisfy many of the IDN rules, including the [inaudible] rule, automatically because of the limited repertoire.

But we have a common set of default rules that will be true for every single LGR in the root. And our current plan is to develop a scheme that we don't have to actually physically repeat those. We've not done that [inaudible] but that's our idea. That's where we want to go.

And I might add we're also not really assuming, and that's partially an answer to the previous question, that people will hand edit the XML files because the way you can do it – the way the Integration Panel did the MSR – the easy way to do it is if you buy the script that takes whatever your favorite plain text format is and then goes through a little pro-script and spits out the XML. XML is not a format that is easy for hand editing. We're not encouraging that.

One of the things that we can do with our own tool implementations that we did for verification is we can take even existing IDN tables and convert them back and forth between that and the XML format. We did that as part of verification, but in the process, we learned that some of the editing work is really easier done in the plain text format that doesn't have any syntax elements in it. You can just write a list of code points and then have a script that goes through and just make up the beautiful XML and doesn't leave out the angle brackets and doesn't leave out closing slashes and does all that perfectly, and all you have to

load your script is a bit of meta data and stuff like that. Then you can do it. Michele did that work for us in the MSR.

KIM DAVIES: My quick take on this is the LGR format is no prescriptive and not tied to IDNA in any way. You can describe code points that are not valid in IDNs at all.

In the domain name context, however, it's essentially a two-pass process. You use LGR, and you do an IDNA round trip. When you're implementing the IDNA round trip, it's going to catch context rules, non-PVALID characters and so on. So it's more of an academic question I think in terms of, yes, we can represent all the context rules in the LGR format. Do you have to? I think practically the answer is no because if you're a registry implementing an LGR, if you generate a string that can't possibly be an IDN because it fails the IDN conversion, then it's like a valid result.

So I think it's a minor implementation detail. I don't think you need to belabor it too much. But one design consideration, just to take out of that, is that the LGR specification, we're talking about if for domain names in the context of the root LGR that we're all here for today. Clearly, it's for domain names. But conceptually the format can be used for other kinds of identifiers. It's not limited to domain names, and therefore it's not constrained to certain IDNA concepts.

NAELA SARRAS: Thanks, Kim. Are there other comments? Certainly I've gained a lot more appreciation for it sitting through this presentation. I think my takeaway as a staff member is we need to stay aware that the work of the Generation Panels is to generate the rules, and whatever we can do to help represent the rules I think we need to take on that work. So whether we do more of these or any other process we can facilitate, I'd like us to continue to be aware of this.

So Marc, did you have something? Go ahead.

MARC BLANCHET: I didn't want to interrupt you, but I wanted to say something, but only if you're done or something.

NAELA SARRAS: Go ahead.

MARC BLANCHETL Okay. I wanted to come back to your question about the XML format. So XML format is actually a very good thing because it's a formal specification. But don't look at XML format in this language as being an obstacle for your work. So if for whatever reason it becomes an obstacle or a problem, please raise the issue with ICANN, and whoever will help you by tools, by whatever needed. This should not be seen in any way as an obstacle. It's just a formal way so we're all on the same page, and it's a formal specification.

Having said that, we're here to help.

AKSHAT JOSHI: Actually, when I asked that, because currently the examples are fairly simple, and maybe once we put all these things in place, it may happen that one may conflict with the other and our interpretation would be like, “Okay, I have broke this rule and this will happen,” and if we keep this like we will send it to the Integration Panel and then they will try to integrate, and if there are some errors into that, then it will come back to us. It will be an unnecessarily cycle.

So looking at it from that point of view, I just wanted to see if interface was one of the possibilities. One more suggestion is that if ICANN has a local tool, like if I integrated some rule and ICANN verified that my end only that it works, this is what we have done in the [inaudible] our local implementation as there was a proper implementation of the engine. It is open, and one can put in a string and see if it becomes valid or not. If it is valid, then engine shows it as valid. If not, it comes a better response saying this kind of deformity, that kind of deformity or not. I think that would minimize the [inaudible] of the LGR.

ASMUS FREYTAG: Right. One of the things we’re communicating as the Integration Panel in particular, and it’s not the P1 project for the tools, but the Integration Panel, but there are certain things about an LGR that pose potentially particular issues in integration, and we always the Generation Panels that are attempting to use those features to be very proactive in contacting us so we can work together to make sure that if there’s conflicts between the Integration Panel and the Generation Panel, that

there should be, only on matters where we have really different opinions on what needs to be done – not just a stupid silly technical mail function. We don't want to hold up the process. We don't want to have the process sit through another public review cycle just because file format didn't work. So we're going to work with you on that.

But I think the issues are sufficiently different panels, like for Latina and Greek and Cyrillic we don't expect to use any features that need difficult review, and Chinese, Japanese and Korean have issues with variants that you don't have.

So at this point, from the Integration Panel, our operating assumption is that will have to not spend our time in trying to figure out generic solutions, but work with existing panels, and if something repeats, then maybe we can improve and make something more generic. But initially, we are on a discovery mission.

Also, Arabic has separate issues. We're open to work with all of you to understand what we can do, and if all of you have the same problem in the same spot, then we go back with Kim and see if we can do tools or whatever.

NAELA SARRAS: So in the interest of time, and we have a fairly large session, so I want to help them facilitate to start on time, so noted what you said, but Asmus, I have that in my notes and we'll work on that some more.

So we have a comment online, right? Or a question. Let's see if we can get that out and wrap up.

UNIDENTIFIED MALE: Yes. One more question for James Mitchell from ARI Registry Services: is the aforementioned test LGR tool publically available – the one that converts rules to regular expressions?

KIM DAVIES: Not at this time, but as they're maturing, our goal is ultimately to have open source tools available.

NAELA SARRAS: Okay, thank you. So I hope this was helpful – this last bit. I urge you to please continue to communicate to us and ask us for more resources, more of these presentations – whatever it takes.

I want to thank everyone for your participation in this whole four hour session. I realize we put a lot of topics, but as staff, we want to thank all the presenters, all of you for attending. This has been very helpful and insightful for us. I appreciate your time and energy being here and working on this project, and we'll see you again in London. But before we have another session here for IDNs here, so we will close it and go to the next session. Thank you.

[END OF TRANSCRIPTION]