LOS ANGELES – IDN Root Zone LGR Workshop
Wednesday, October 15, 2014 – 10:00 to 11:15
ICANN – Los Angeles, USA

SARMAD HUSSAIN: So let's get started for the next session. Could I request everybody to please settle down?

So this session is focused on IDN root zone LGR. This is more from a community's perspective. We have community presentations in this session.

Instead of just community updates, the focus of this session is to look at a different aspect of what is needed as generation panels do their work. The aspect we will be discussing or focusing on today is actually the coordination aspect of different scripts.

I'll just give you a brief introduction to how are we looking at different coordination scenarios. Then we'll have Marc who will give you a really brief overview on the [inaudible] limitations and mechanisms for the root zone LGR.

After that, we will go into community challenges by Meikal Mumin on Arabic coordination across multiple languages. Then Wang Wei who will talk about coordination challenges Chinese, Japanese, and Korean scripts and how the community generation panels are regressing those coordination issues.

Then we have very equally interesting coordination challenge and possible solutions being considered for Neo-Brahmi scripts. The presentation will be done remotely by Nishit Jain.

And finally, the [possible] coordination requirements between Cyrillic, Greek, and Latin scripts by Cary Karp before we go into a question and answer session at the end.

As an overview leading into this discussion, basically we are witnessing multiple scenarios here as we formulate the generation panels, and therefore different kinds of coordination models need to come out, and those coordination models need to really come out from the communities themselves.

However, there are four different phenomenon we are looking at. We have cases where you have many languages [inaudible] using single script – one script – and it's dealt with one generation panel. Arabic is an example for that.

We also have one script being used – being worked by many different generation panels. Example is Han script being used by Chinese, potentially Korean and Japanese generation panels.

We also have the example of many scripts – different scripts, but all being dealt by one generation panel. We have, for example, [inaudible] Tamil, multiple scripts which Neo-Brahmi generation panel is looking at.

And then of course the final possible coordination scenario is you have many scripts and many generation panels, and the obvious case for that is Cyrillic, Greek, and Latin.

So we will actually have presentations from all these different script scenarios and share what coordination challenges are and how these communities are considering working with each other.

We obviously, when we're looking at coordination, we want to look at two aspects. The need of coordination, meaning what collaboration needs to be done. Is the need only based at code points or the need of coordination is also for variance? Is there need to coordinate across different rules or are there other dimensions to coordinate when we are talking about multiple generation panels?

And then the mechanism itself. For example, the question to ask is should the coordination be done before the generation panel does their work, or should the coordination be done while the generation panel is doing its work? It should also be coordinating with other generation panels. Or basically the generation panel should really finish their work first and then go into a coordination activity.

Each of these models have their own practical limitations and advantages. Obviously doing coordination before helps you plan, but it stalls the whole progress. Whereas doing it afterwards is most optimal as far as internal working is concerned, but then you may actually have to do some pre-work if you go and do the coordination [inaudible].

So what is the best way of actually coordinating where you can balance – you don't have to do a lot of [rework] but you can also make progress internally as well.

So those are some of the things we will be looking at, and without further delay, let me pass on to my colleague, Marc Blanchet, on behalf of Integration Panel who will just give an overview of [inaudible] limitations and mechanisms for the root zone LGR. Just a broad introduction. Marc?

MARC BLANCHET: Thank you, Sarmad. I'll go, not fast, but some of those topics have been already discussed in some ways. So since this process has started, sometime we were discussing with possible generation panels and community. There were a lot of discussions about variance, which are an important part of this. But the LGR is first identifying the set of code points of the repertoire. So really the need for LGR is not only not all about variance.

The LGR really defines what labels are valid for automation, and for some scripts, all is needed is just a set of code points. Obviously the root repertoire itself is a collection of single script repertoires that are each tagged by script. There's no cross-repertoire labels and no overlap except common code points, which is [inaudible] case.

I said before each script repertoire is limited to modern use, everyday, and stable.

So what about variance? Some scripts require variance, [inaudible] definition of variance which is not confusable, but the same for the user's point of view.

There is two types of variants that could be – mechanisms or dispositions that are blocked or allocatable. And the procedure, actually try to maximize the number of blocked variants and minimize the number of allocatable variants for scalability and various other security concerns.

Variant mappings will be used to automatically generate all permutations. So therefore, the more variants you have, the more permutations will become and they may be a [large set].

So the type of variant mapping determines whether the variant label will be blocked or will be allocated to the same applicant. As a result of integration, all the blocked variants can exist across generation panel repertoires. So generation panel coordination will ensure the consistent outcome of this process.

Whole label evaluation rules. Why this is important to understand that they are needed to prevent labels that cannot be processed or rendered. This is not about importing the grammar or the syntax or autography of a language or script into the LGR. This is not about that. IT's really about specific needs of labels that cannot be processed or rendered. It's generally for complex scripts and may not be applied to your scripts. So this is a place where there's, for some scripts, no need to carry or take care of WLE and some others will.

So as we probably all know, the TLDs are intended for good mnemonic value and not intended to capture and implement all the facets of a writing system.

What should be coordinated? In the case of repertoire is consistent with treatment of similar repertoires. An example of this is the Indic scripts. Variants is a compatible definition of variants. So the disposition can be different, but the actual set of variants must be consistent. And whole label generation [inaudible] consistent treatment processing of structurally similar scripts.

I would like to conclude with one thing which is the bottom URL here which says "Variant Rules." This is a very good document that descripts in depth all the information about variants. They need to be defined and stuff. If you care about variant rules, I really encourage you to look at the bottom URL, which is a document that was produced by an integration panel that describes a lot about variant rules. Thank you.

SARMAD HUSSAIN: Thank you, Marc. So now we will go into generation panel – community based generation panel talks. We will start with Meikal Mumin who will be talking about the work challenges in addressing multiple languages using Arabic script. So over to Meikal.

MEIKAL MUMIN: Yes. Hello again. I'm still Meikal Mumin. I'm still here to represent the Arabic Generation Panel. I'll try to give a summary of the challenges in addressing multiple languages using Arabic script.

So what we are doing is we are representing scripts in the world of languages. So on the left-hand side, you can see Roman/Latin script IDN, and on the right-hand side, you can see an Arabic script IDN. But we don't know which languages are used by websites of either IDNs.

So international domain names do have a script as property, but not a language. So what does it mean? Well, it means that IDNs cannot be based on the orthography of one language, such Arabic language, but that the LGR and related standards must therefore address the entire communities of readers and writers of Arabic script.

The problem is that, while we can only represent scripts, we think in terms of language. So all data is at language level while have to define LGR at script level. There are no institutions representing script communities and writing is usually considered a reduced representation of language.

So what is the actual scope of Arabic script LGR? Well, Arabic script obviously is centered around Africa and the Middle East as a writing system, but in the course of time, it has expanded across nearly all continents with established past or present use in the Americas, Europe, Asia, and Africa.

And only within Africa there is attested past or present use of Arabic script for the writing of more than 80 languages apart from Arabic.

With today's patterns of migrations, continuing proselytization, and population growth, we have to consider the fact that more user communities of Arabic script are manifesting in both the Global South and North.

Accordingly, Arabic script is used not just locally or regionally, but globally, albeit to radically different degrees and entirely different manners, since for numerous languages, Arabic script is an active competition with other scripts, it is only used [inaudible] by only a part of the language community, and it's really not foreseeable how the situation will evolve in the future and what the impact of IDNs would be in the community.

So to give you a more extreme example, would a community possibly care of they can register a domain name using the orthography of the

language if any reading and writing is really only conducted with a pen and paper?

So what we have to do is represent the underrepresented because linguistic diversity I was explaining earlier unfortunately is not well-represented, so there is a lack of data on languages and orthographies. There is a particular lack for languages of low status or socio-economic participation and there is little available generally non-western orthographies while non-standardized orthographies are generally not considered.

So often much TF-AIDN has to rely on user intuitions from an entirely different part of the script community. One problem we had, for example, is during code-point analysis. We frequently lack data to establish whether a code point is used optionally or obligatory in given orthography, which is required within the current process.

So one of the things we asked, according to the current standard of MSR and the label generation rules process is to quantify and qualify the script use according to the EGIDS scale which was discussed earlier.

Since security and stability of the DNS and root zone are highly important, conservatism is a strong principle, and as a result of that, the integration panel has said in the Overview and Rationale document that "Where the Integration Panel was able to establish to its satisfaction that a given code point was assigned a character solely for use in a disused orthography or for a language in serious decline, the code point has been removed from MSR."

Now, MSR dictates that the Expanded Graded Intergenerational Disruption Scale is used to categorize, as they call it "effective demand of languages within a given [community]."

And this scale principally consists of 13 levels which rank languages from the highest representation in [inaudible] society being a national language to the lowest, which is extinction.

So for the MSR, the IP, as mentioned earlier, uses a cut-off which is between levels 4 and 5. Unfortunately, we have to accept that such representation of language in society is not just accidental, but usually a result of historical processes.

So Andy Warren-Rothlin recently said that "Scripts divide languages into cultures, make dialects into new distinct languages and create new dialects. If, as it is often, a language is a dialect with an army and a navy, how much more is it a dialect with a distinct script?"

So we have to be aware of the fact that people, society, and language play a role for IDNs, because languages and scripts are evaluated by people, are assigned a status by both societies and scientists, and are regulated by governments. And this is reflected also in studies and statistics on languages.

So to give you a more extreme example, there have been historical reports of orthography suppression of Arabic script where the use of writing systems actually has been banned and criminalized. Obviously, in such contexts, it will be very difficult to get qualified data on the use of that writing system.

So my point here is simply that I guess we must be cautious not to strengthen further trends of linguistic discrimination and strive for equal treatment of languages, even where they lack socio-economic participation or political representation, and as a proxy of that, representation and the data accessible to us.

So, for example, TF-AIDN did identify some 32 code points during their analysis of available code points with evidence of use which cannot be included in LGR because they do not have the necessary rating, according to that scale.

Let me provide you with two examples. One is, for example, the language Seraiki, which is the language of Pakistan. There are, as a matter of fact, numerous publications in Seraiki including daily newspapers. But within Pakistan, Seraiki has a rating of 5 and IP recommends a cut-off of any language with a rating lower than 4.

Also, we have a language called Harari, which is the language of Ethiopia, which within the Ethiopian context is comparatively small language community, but there are significant expatriate communities which seem to be very active.

So, for example, there are associations in Australia which have published an orthography description and a virtual keyboard with the assistance of the State Library of Victoria, Australia.

So within Ethiopia, Harari has a rating of 6a, so below the threshold. But it has no status in Australia in the expatriate communities.

So because of the activity of the expatriate community, TF-AIDN assumes an active use of the orthography and would suggest such

relevant code points used by these orthographies for inclusion. Unfortunately, this is not possible within the current positive stipulated.

So, to some degree, we are dealing with principles which are laid out before TF-AIDN and the Arabic script label generation panel could conduct its analysis, because if we look at the process, in the case of Arabic script IDNs, ICANN has tasked two groups to work together to develop those rules.

Once the Integration Panel who published also the Maximum Starting Repertoire, and on the basis of the procedure laid out, as I was informed by my colleague [inaudible], it's not in fact the Integration Panel but another Board of ICANN which published the procedure, and on the basis of the MSR published by the Integration Panel itself, TF-AIDN should formulate the LGR which is then approved by the IP.

Accordingly, the rules have been laid out before we could conduct our analysis. So in a sense, like I said, the development process and the rules have been designed before an ideally data driven code point analysis could be conducted by script generation panels.

The Arabic script generation panel and TF-AIDN possibly is only the first to notice this because we are the first script generation panel to take up work.

On the basis of that, we did suggest as public comment to Integration Panel that MSR-1 should only be frozen one script at a time, and that after relevant script generations had to have the chance to give their feedback on the relevant portions. Unfortunately, this was considered removal of MSR-1, and therefore refused.

So let me give you a last example of issues we're dealing with, which is variants which were just mentioned. Again, they are required to balance the usability of IDNs as well as the representation of languages against security and stability of DNS and the root zone [inaudible].

And there was an Arabic Case Study Team, as described by our colleague, Sarmad Hussain, earlier who published a report in an earlier stage before formation of IP and before formation of the MSR, which identified six types of variants in Arabic script.

So here I can give you two examples. On the left-hand side, we have one example from a group which is considered variants. On the right-hand side, you have another two examples from a group which is considered a type of variant.

I'm not sure if you can read it from here. Unfortunately, it's smaller than is on my screen. But the point is, in a sense, these letters are similar, yes. But on the left-hand side, it is the basic letter shape which is similar, while the dotting pattern which is part of Arabic script is identical. And on the right-hand side, the basic letter shape is identical, but the dotting patterns are similar because they are rotated.

Now, we're supposed to reasonably argue that the difference on the left-hand side in letter shape is not confusable by all readers and across all representations and fonts, because this is what we have to base variants on. It must be across fonts and across users, across the entire script-using community in a sense.

While this difference is not confusable, as is currently the trend and analysis within [inaudible] internally, that the examples on the right-

hand side are not confusable to at least a subset of readers or in a subset of representations or fonts.

But the fact is there are no empiric scientific tests to support either theory, and obviously there is a systemic bias in representation even within their own group, because 15 out of 29 members are first language speakers of Arabic.

So my point is if the majority of people who judge on what is confusable and what not comes from one language community, it's probable that also the analysis could be bias. But it's just an issue we have to bear in mind. Anyways, thank you for your kind attention.

SARMAD HUSSAIN: Thank you, Meikel. We'll just keep moving on and we'll take comments and questions at the end of the session. So for the next presentation, I'll request Wang Wei to give a quick update and also focus on the challenges and solutions which are being undertaken for coordinating between Chinese, Japanese, and Korean script generation panels. So over to Wang Wei.

WANG WEI: Thanks, Sarmad. The coordination – we're doing the coordination between script panels just like Marc mentioned. The [inaudible] here is we have one script, but we have three different language scenario.

So what we're trying to do is to avoid the conflicts that happens because of the overlapping code points between three panels, and

because the fact that the Chinese code points are used in different ways in three different language [environments]

This graph is trying to elaborate the overlapping relationship between three panels. The Chinese character repertoire or Chinese code points exist in different scripts.

So far, the Chinese character table and the repertoire rules under RFC 3743 and 4713 has been widely accepted by the .CN, .TW and .HK in Chinese area and Chinese language area. And for Japanese, the second-level domain table under JPRS, under .JP, there are over 6m000 Chinese code points and overlapping code points is 6,186. And for Kisa, for .KR, so far the Chinese code points are not allowed on second-level domain registration.

So we have different variant solutions in different scripts. For Chinese script, well, we will allocate the applied IDL and variant IDL to the same applicant, but only the original applied IDL and [inaudible] simplified and traditional IDL will be delegated. The others will be reserved – or what do we say? Blocked.

For Japanese, there is no variants. Whatever is old form or new form code points, they are treated as independent code points.

And for Kisa, there's no Chinese code point, so there's no variant issues at all. I say so far, because Korean community hasn't decided whether they should allow the Chinese code point on the TLD level for the root.

Marc mentioned the coordination principle. One key principle is that the variant mapping must agree for the same code points for all LGRs. I think a compatible definition of variants is necessary, but during a

discussion with different panels, we think if there are some words to elaborate the variant mapping or the variants, to elaborate that terminology, it will be better for different panels to have a better understanding.

For example, for Chinese code points, the variant mapping, maybe it means code points family with possible variant mapping relationship in different panels. It would be better if we had some elaborating words that will help different panels have a consensus.

Also, we have several case studies to help different panels to get us to [inaudible] consensus. We have three different cases. The first case is that, for Chinese, this code point family has – or variant family – has two code points. But in JPRS second level repertoire, there [is] only one.

So if Japan could – [inaudible] Japan to add the Chinese code point into their repertoire and block this Chinese code point. It doesn't exist in the current Japanese repertoire.

Another case is in CGP repertoire, the variant family has [four] code points which exist in the Japanese repertoire. In that case, they just exist parallel in different panels. The Chinese panel has their own rules and the Japanese have their rules.

Another case is that for Chinese panel, the variant family has two code points. But in JPRS table, there were three, [even though] they are not treated as variants.

In that case, considering the fact that the third code points in JPRS table is [inaudible] Chinese code point which were treated as a variant of the

above two code points, the CGP would add that code point into the current table and block this, abolish the Chinese engine code point.

The final case is that there were some Japanese unique characters which are not treated as Chinese code points. In that case, we are not trying to test this code point and just leave it in the Japanese table.

The expectation for JGP and KGP? Well, for JGP, Kanji which means the Chinese code points in Japanese language, repertoire and the variant type annotation are expected. For KGP, it is important for them to decide whether or not allow Hanja which means allow the Chinese code points for Korean language. If Hanja is allowed, then its repertoire and variant annotation is expected.

We should work together on trying to reach a unified variant mapping table for the overlapped code points. We already have four cases, but maybe there are more scenarios when Korean community published their repertoire. The situation may be more complicated.

The challenges? The first challenge is that CDP is trying to finish this job by the end of this year, but maybe this work plan will be extended for the synchronization between C, J and K.

The second one is repertoire modification. As I mentioned, there were four cases already. We have to analyze the code points overlapping issue and for different scenarios, then [negotiating] among the panels is needed and we need to modify our current repertoire and the current variant type annotation.

The final one is the whole label generation rule set. Each panel should be aware of the pros and cons of the language tag based solution. Of

course we should focus on the technical issues, but the language tag based solution will influence the future registration.

If the variants definition which means – well, if there's not a clear variant definition of the variant mapping definition, it will cause different understanding of the future registration rules. Some panels are worried about that if the whole variant package were to go the same applicant or not, [it would] definitely influence their current decision about this generation rule.

Thanks. That's all.

SARMAD HUSSAIN: Thank you, Wang Wei. We'll keep moving ahead in the interest of time. The next presentation on Neo-Brahmi script by Nishit Jain. He will be presenting remotely, so could we have him on the speaker?

NISHIT JAIN: Thank you, Sarmad. Am I audible?

SARMAD HUSSAIN: Yes, we can hear you. Please go ahead.

NISHIT JAIN: Thank you very much. My name is Nishit Jain. I'm from [inaudible] India. [inaudible] organization of the Department of Electronics and Information Technology [inaudible] IT and associated areas.

Our group [inaudible] is mostly involved with [inaudible] language processing, especially in Indian languages. I'm a member of Neo-Brahmi Generation Panel. Today here along with me I have [inaudible] who is also a member of Neo-Brahmi Generation Panel. This is a brief introduction about me. Now I will start with my presentation.

In my presentation, I will talk about Brahmi script, because initially, [there are] variant issues project was started. We were talking about the [inaudible] script only. But here, we are taking Brahmi script into picture.

Then we'll [inaudible] about Neo-Brahmi and the current status as well as the outreach effort of Neo-Brahmi Generation Panel.

Further, I will talk about the issues which may arise and the approach which Neo-Brahmi Generation Panel may follow. So let's start.

Brahmi, [inaudible] script which was [inaudible]. It is an old script and not in use today, but it is the ancestor of most of the modern [inaudible] IDN scripts, like [inaudible]. So Brahmi acts as a mother of all these modern user scripts.

Geographically speaking, these scripts are being used in Central Asia, South Asia, and Southeast Asia. There are many languages which are used in these scripts, especially the languages of Indo-Aryan family and the languages of Dravidian family. Next slide, please.

So despite the visual differences across these scripts, the basic foundation or the basic philosophy between them is common. That is they all are akshar driven. An akshar is the basic processing unit or the

fundamental unit which is [inaudible] recognized by the native user of the language using Brahmi based script.

So digital medium, an akshar is represented using a syntax. This syntax plays a vital role in whole label generation procedure of IDN. Next slide, please.

Why Neo-Brahmi? Because out of all the scripts or families which are derived from Brahmi script, we are not [considering] all of them, but we are taking those scripts which are in the modern use today. This approach is in consonance with the "Conservatism Principle" of LGR procedure. So this is why it is Neo-Brahmi Generation Panel. Next slide, please.

So there is some similar work which we have done previously. We are working on [government affiliate] project for IDN ccTLDs for 22 official languages of India, which is IDN version of [inaudible] ccTLD. And under this project, we have carried out similar activities.

So for each language, we have finalized a set of code points which are permissible. And the visually similar variant strings are also called [homoglyphs]. The whole label evaluation rules, which are complex in nature and adheres to the principle of [inaudible].

So these language policies were framed based on the input received from language experts as well as from community representatives, and also from the feedback received from general public.

Recently we have launched .bharat ccTLD in Devanagrai for eight languages: Hindi, Marathi, Konkani, Boro, Dogri, Maithili, Nepali, and Sindhi. Next slide, please.

In context of LGRs, we might have to revisit these rules which we have laid down for the .bharat ccTLD. The policies under .bharat ccTLD were framed [inaudible] in mind. But in the case of LGRs which are for the root zone, the stakeholder group is much wider. The comple whole level evaluation rules might need some revision to comply with the principle of simplicity and predictability laid out in LGR procedure.

However, this revision would not change the basic principle of Akshar formalism for well-formedness of label. Next slide, please.

So the Neo-Brahmi Generation Panel currently consists of 10 members who come from different backgrounds of linguistic and Unicode. We are still in the process of getting more members on board. I hope we will get enough participation by mid of November. Then we would be able to send Neo-Brahmi GP proposal to ICANN. Next slide, please.

As an outreach effort, we have conducted our workshop in AprIGF-2014 as an awareness program and call for participation in Neo-Brahmi GP for the formation of label generation rules. The topic of presentation, the workshop was “Bringing diverse linguistic communities together for a unified IDN ruleset” and [inaudible] wide area of discipline had participated in that workshop to provide a holistic overview of the topic.

In the workshop, various aspects of creation of LGRs for Neo-Brahmi scripts have been touched upon.

We have also given updates about Neo-Brahmi Generation Panel in previous ICANN meetings. The main objective of all these workshops and presentations was to reach out to the community for wider participation. Next slide, please.

Within these modern derived Brahmi scripts, there are many cases of characters being confused about with the characters of other scripts. We can characterize them into two different classes of character confusability.

The first one is [inaudible] confusable which are both close [homoglyphs] as well as homophones. The third row of the table shows an example of two confusables where the Devanagari character is both [homogloph] as well as homophone with Gujarati character.

The other category would be [inaudible] confusable, which are close [homoglyphs] but not homophones. First two rows shows the example of [inaudible] confusable. This part is given in a more elaborate way under section 3.5 and section 4.1 of the [inaudible] Variant Issues Project Report, which is published on ICANN website.

However, we anticipate the script mixing within our TLD would mostly be [inaudible]. Allowing that would have caused a bigger [inaudible] which is possible by mixing of these characters from different scripts.

So even after that, use of only these characters might give rise to similar looking strings across scripts. For example, I show [inaudible] in the cases of similar looking strings between Devanagari-Gujarati and Devanagari-Gurumukhi. So the Neo-Brahmi Generation Panel would definitely want to consider the issues of cross script similarity. Next slide, please.

Since Neo-Brahmi Generation Panel is intended to cover modern scripts which are derived from Brahmi script, we would prefer to organize our generation panel into sub-groups.

Initially, we will be focusing [inaudible] gTLDs or ccTLDs are applied or delegated. So the Neo-Brahmi Generation Panel would consist of a chair, the policy experts (Unicode/IDNA Experts, Registry/Registrar Experts). Then there will be sub-groups specific to what script which consists of the language experts as well as the community representatives.

So the policy expert, Unicode experts, registry/registrar experts will coordinate with each sub-group for the formation of LGR for the respective script. Next slide, please.

There are cases of one script, one language and one script, multiple language. Under [inaudible] generation panel, multiple sub-groups will coexist. However, one sub-group is intended to handle one script. But in case of multiple languages are using the single script requires a [inaudible] of diverse linguistic experts to be there and [inaudible] sub-group so that each language can be represented in a proper way. Next slide, please.

This is an internal composition of Neo-Brahmi Generation Panel. This is what I was referring to in my previous slide of scripts like Devanagari being used by multiple languages [inaudible] requires a diverse linguistic expert to be there or [inaudible] sub-group to represent each language in a proper way.

I think that's all from my side. Thank you very much.

SARMAD HUSSAIN: Thank you, Nishit, for the overview and the structure of coordination being undertaken for Neo-Brahmi script. And now we move on to our

last presentation for the session on the coordination between Cyrillic, Greek, and Latin scripts – an overview being given by Cary Karp.

CARY KARP: Is somebody going to click for me? Okay, great. First slide, please.

Here again, because Latin is a part of this, we're building on something that's already there. It's not a part of the IDN exercise. So this is going to be lighter weight than the others need to be.

I don't know if that's actually a word or a joke, but if it is a word, it's actually a pretty good one, a pretty useful one. The letters, the characters Apsyeoxic exist in identical glyphs in both the Cyrillic and the Latin script. It's probably the longest phonetically coherent thing you can concatenate using them. But one way or the other…Next slide, please.

If we take a look at the glyphs associated with code points, code from the MSR for both Latin and Cyrillic, there's a fairly-reasonable degree of overlap there. Those are absolutely congruent glyphs, too. It's not a question of visual similarity. It's a question of identical bitmaps.

So the Latin A, which is at one code point, is commonly displayed exactly as is the Cyrillic A at a completely different code point. And there is no code point overlap or confusion. It is the glyphs that we're talking about – the bitmaps.

I don't know if that's a large enough array to be a cause of concern and require coordination between the Cyrillic and the Latin groups. It's just a

question. I'm noting the problems that we might have. If we go to the next slide…

And include Greek. Now we do need to inject a little bit of latitude for similar glyphs – if not absolutely identical –and I've sort of lined them up vertically. There is – you can see what the overlap is there. And we not that this is less than the overlap between Latin and Cyrillic. And a comparison between Greek and Cyrillic, it's the same deal. That if we're talking about the three, generally regarded as very closely related alphabets – certainly the two alphabets that are most proximal to the Latin script, again which has no linguistic privilege, but it is the thing on which the DNS is based. The script used for domain name labels prior to IDNA.

We can go a step further. Now, this is beyond the horizon of IDNA. It's beyond the horizon of MSR. But it is not beyond the horizon of the concerns of the user community.

And we note here – if we look at the upper-case letters, Cyrillic, Greek, and Latin, the degree of overlap becomes significant. The Latin community would likely say, "Hey, wait a second, do you mean we are not going to be able to use mixed cases in IDNs?" And the answer to this is that's absolutely right. BigDeal.com can be written capital "B" small "i" small "g" capital "D" small "e" small "a" small "l" and things will appear to work just fine.

But if you decide to put an accent on any of those small letters, all of a sudden everything collapses. It has to be all lower case. And there's a trade-off there. From the perspective of the Latin community, do we

want the camel-casing or do we want the decorated letters? Again, a question.

As I understand it, on the basis of the result of a similar discussion conducted during the course of the VIP studies where representatives for each of these communities met, from the Greek perspective, Greek users simply expect to be able to enter upper-case characters and anything they do with a computer and there is no way they are going to be taught to do otherwise, which means that the availability of IDN is a matter of some concern if on no other level than that.

Again, I have no idea how that maps into what we are doing other than it presents itself as a concern of the communities that we are serving. And we can be Draconian about it and say we're really sorry, but there's nothing that can be done about that – which is in fact the response – but there may be some degree of need for a tutorial statement or something. Again, it's not just quite as convenient an algorithmic exercise as is difficult enough already. It's worse than that.

And again, IDNA deals with it simply by saying if you need to address issues of case, you do it before you send anything to an IDN compliant engine. But that doesn't mean that these issues are irrelevant. Next slide.

It simply means that the LGR panels, at least for these three scripts, presumably for any of the overlapping ones, need to discuss what is in scope and what is out of scope for their deliberations with a very convenient given response that, well, neither IDNA or MSR permits any consideration of case so we're not going to consider case [inaudible].

So I would hope and I will yield a moment or two of my time to discussion, given that are all the communities here? I mean, you were in the room at the time when our Greek colleagues that we have a huge problem. And you were interested in that problem. I was busy covering my eyes and ears. I didn't want to know anything about the problem. But now we're there and we're there big time. Otherwise, there's nothing more I have to say about this.

Again, otherwise I'm beyond what I would regard is important to stress. The dictates of the reference documentation for this exercise provide extreme focus on what we're doing, but the expectations and the need for explanation to the target community for this entire effort goes beyond that. We don't just need to do something. We need to have it understood.

Okay, thank you.

SARMAD HUSSAIN: Thank you very much, Cary, for an excellent presentation on potential issues and overlap between Cyrillic, Latin, and Greek scripts. Before I open the floor to comments and questions, I request Asmus to make a quick comment on what integration panels' expectations are [inaudible] required between certain generation panels for the various cases we have seen. And after that, we'll start taking questions. Asmus?

ASMUS FREYTAG: Thank you, Sarmad. I appreciated the very in-depth analysis of the coordination issue that was presented in some of the presentations – in particular, [inaudible] look at the Latin, Greek and Cyrillic.

The Integration Panel expects the result of coordination essentially to make its job easier in that the related scripts and the communities associated with [them] have come to a common and workable understanding of how to treat issues that arise from related scripts, so that the Integration Panel is not left stuck with having incompatible LGRs and potentially has to reject one or more of them. We are not in the business of wanting to reject LGR proposals.

In some cases, I think it is entirely appropriate for Generation Panels to think through some of the related issues like the effect of user agents presenting URLs in upper-case when the underlying IDNA are required to be only lower-case. I think that it's entirely appropriate to look at those issues even if they don't have a direct representation in the tables that are ultimately being created and delivered.

Looking a little bit over the edge of your plate to find out how your solution is going to impact things on a larger scheme is going to make it more robust. There's nothing wrong with that.

The overall hope is that the Integration Panel has is when there's a need to define variants or define blocked variants across repertoires, across scripts, across panels that we get a consistent model for related script presented to us and can proceed from there.

We are definitely very interested in engaging a new generation panel that is undertaking that work early to have discussions so we can understand the issue, we can understand where they come from, there are no surprises in the process and if we have Integration Panel concerns with some particular projects, we would like to take those up as early as possible in an informal stage so that we don't have to have

unfortunate realities when we have a formal review of final proposals. At that point I think we would love to have all issues already resolved so we can just go ahead with integration.

SARMAD HUSSAIN: Thank you, Asmus. And for possible interactions with Integration Panel, ICANN staff can provide opportunity to the generation panels at obviously the ICANN meetings, but also through conference calls. So if there are needs or concerns, that kind of dialogue can also be arranged.

Before we move on to questions, there's a few comments online through Adobe Connect. I'll ask [inaudible] to read those comments.

UNIDENTIFIED FEMALE: Michelle [inaudible], IP member. "IP does not freeze script context. Further version of MSR can augment script content if needed."

Second comment, [inaudible]. "I'll explain GP status. RFC 5895 provides previous mapping for IDN strings."

SARMAD HUSSAIN: Okay. We can now open the floor for questions or comments. Does absolutely have any questions or comments? [Umer]?

[UMER ANSADI]: Thank you, [inaudible]. I really enjoyed the discussion today. My name is [Umer Ansadi]. I'm [running] a tech firm based in Afghanistan. Before that, I was working with Dr. Sarmad Hussain under a panel localization program that he was leading on software localization. A part of the

activity included localizing IDNs into Pashto, which is an official language of Afghanistan.

But the thing is, the localization we did, it was never implemented. That was back in 2007. And we would like to restart the process. After the panel localization, Dr. Sarmad, we never had a chance to work on building upon the activities we did under the Panel Localization Program. What is the process for the new communities languages that are not included underserved to get into the process, and what are the roles of various communities in groups including the technology community, local governments, ICANN? And whom we should be speaking with if we would like to start?

SARMAD HUSSAIN: Just to respond to that question, if you want to participate in a generation panel – and in this context, it would be the Arabic Generation Panel – we can provide you with the information. Arabic Generation Panel is open, which means that it takes members at any time. It is still open and it is still taking members for those who want to contribute. So I will share the e-mail address [inaudible] which you can share your interest. I guess you can start contributing to the process.

There was another question. [Vladimer]?

[VLADIMER]: Yes, two questions. First, relating to coordination between Cyrillic, Latin and Greek. I don't think anybody mentioned the completely different mechanisms that that exists for the root TLDs which is string similarity and panel. I assume it was designed there to come and stay, meaning

that it's pretty much guaranteed that in any new TLD program, before a decision is made to allocate specific top-level domain it is going to pass through string similarity evaluation.

That brings up the question. So if the problem of similarity between – if the need of coordination between Latin and Cyrillic is motivated by the fact that there are similar characters, isn't that problem already addressed through the similarity evaluation mechanism? If not, could you give an example where we need to deal with it here within the variant TLD process rather than string similarity process?

ASMUS FREYTAG: If I may answer to that one. The problem is overlapping. However, most of the examples that Cary provided, it is not so much a question of the characters being somehow similar to each other. It is a question of that these are fundamentally the same entities that have historically been used in different context while still being recognizable, the same entity, and have been given by Unicode separate code points for that reason.

However, that decision by Unicode to give separate code points is itself, in some sense, arbitrary. You can prove that if you want by the fact that there exist or existed encoding schemes that did not make that separation, but provided a unified encoding for those.

To a large extent, it is the opinion of the integration panel that whenever there is this kind of underlying identity with dual coding that that is an issue that it needs to be at least evaluated in this level of the process and potentially addressed by the creation of blocking variant rules. It is not something this process can just kick upstairs to some

similarity review, because fundamentally, the driving issue is not similarity. The driving issue is the underlying identity of the object that has been encoded multiple times.

Just the fact that something – and you can see the underlying identity in the case of Latin, Greek, Cyrillic quite clearly by the fact that the subset that I'm thinking of are try homoglyphs. They are precisely identical in all fonts in all environments. In fact, font suppliers will just simply have a single shape and just map it multiple times internally.

So you really have something that is fundamentally intentionally depicted the same way and not accidentally similar.

Because of its systematic nature, I personally think there's a certain benefit of addressing it early in the stream of cascading evaluations and not dragging a process like the similarity review which has a large overhead of having to look at labels in [inaudible] with appeals process and also [inaudible] when you can potentially get the generation panels to agree that these things are systematically the same thing and should be vigorously treated that way.

I'm encouraging the panels that are forming for Latin and Cyrillic, and hopefully also Greek to look at these issues, talk to each other, see whether there can be an agreement reached of systematically treating these, and then presenting their agreement to the Integration Panel. That would be my hope, that we can proceed in that way.

ASMUS FREYTAG: Thank you. There was a second question. Completely different topic. So the LGR mechanism can have two outcomes, which is the variant is

blocked or the variant is allocatable. Can somebody please give an example, give a case, where a variant should be blocked or the variant should be allocatable?

UNIDENTIFIED MALE: Here's a case. I can't give you one for Cyrillic, but I can give you one. In Arabic, there's two different forms of a certain letter. One is used in the person environment, one is used in the Arabic environment. Users in one environment have one on the keyboard. Users in the other environment have the other on the keyboard. If you have a global system like the root, you want users from both environments to be able to reach the same domain no matter which keyboard they have.

In that case, making those two code points allocatable variants of each other would appear to provide a large benefit to the user community, because whatever keyboard you have, you can just type in what you see and you get to the same domain – assuming that the owner of the top-level domain applied for both forms, because it's not automatic. So that would be an example of an alloctable one.

And the example of a blocked one is given when it's not necessarily beneficial for people to reach the same thing no matter where they come from but to prevent things from existing that appear to be the same thing but go to different sites. So if you take .CO, in Latin goes to Colombia or whatever. If you did something that looked precisely like .CO and had it in Cyrillic or Greek or go somewhere else, that would best be handled by a blocked variant, so that that cannot happen.

There's no benefit of being able to reach Colombia via Cyrillic. So that would be – and that environment may be a potential candidate for a blocked variant.

[VLADMIR]: So the presumption is to block rather than to make it allocatable, right? So in case we are hesitant, we rather block them rather than make them allocatable?

ASMUS FREYTAG: Certainly I think that is a way – that is a certain way of looking at it. Hesitant or whatever. There is a downside to allocatable variants, because on each level, you have an allocatable variant. You have a multiplicity of labels that are supposedly going to the same domain, and if you go down the tree and have many, many different levels and you repeat that process, you can get a [inaudible] explosion of potential labels leading you to the same site.

That is one of the items – I mentioned in my presentation the complexity cost. A very steep complexity cost of having allocatable variants. So you want to be really, really, really conservative and only use them when you're absolutely sure that is required for your script, that the user community really benefits, weigh in excess of the cost incurred by the complexity.

The calculation for blocked variants is much simpler. The main cost is that certain – the second applicant coming in with a label that is equivalent to an already-existing label except for the presence of one or two of the blocked variants in the stings would have their application

denied. This is not much different from a situation where somebody else already has applied for what they desire. So in that sense, the overall cost, the overall complexity is much lower and you can be much more free.

There is certainly an opinion that, in some situations, having the name space and the root a big sparcer by the existence of blocked variants can in fact make the whole name space more robust.

SARMAD HUSSAIN: We are now running out of time. We'll take two questions/comments. [Yuri] and then Wang Wei.

[YURI]: Very quick comment and very simple example. ICANN delegated country code top-level domain in Cyrillic [inaudible]. Okay, [YKP] must be blocked. [inaudible] in Cyrillic, YKP, is a similar case in [Latin]. Must be blocked.

SARMAD HUSSAIN: Wang Wei?

WANG WEI: Yes. I think in Marc Blanchet's slides it is recommended that the definition of variants should be compatible. To my understanding, the meaning of the variants is different for different scripts. For example, for Chinese, the case [inaudible] we have one script but we have three

different language code points we use in different ways in different language environments.

And yesterday I discussed with Professor [Kim] from Korea. I want to ask if it is okay if we propose some words to elaborate the variant mapping in Chinese case. We'd like to propose [inaudible] to IP and if IP can confirm that [inaudible] elaborate the situation in Chinese [words], that would help us.

SARMAD HUSSAIN: So we can take that on e-mail, I guess. Is that something you'd like to do or you want to address that right now? You want to take that offline? Is that okay?

WANG WEI: Offline, yeah.

SARMAD HUSSAIN: Okay. We have some comments online, so we'll go there. We then need to close the session as well. I can see [Edmon] wanting to make a quick comment. [inaudible], before you go, I think [Edmon] probably wants to add something in this context.

[EDMON]: Just a quick comment back on the Cyrillic situation. So I guess it's probably valuable for generation panels, especially Latin area, to think about – because there is also the "protection" on the confusingly-similar string. I know there's an overlap between the two. What the Generation

Panel probably should think about is whether to create a variant relationship or whether it's good enough for confusingly similar to block it, because both ccTLDs and gTLD processes require string confusion kind of process, which would already – current process already disallows that.

So probably the Generation Panel should think about the balance between whether to consider all these blocked variants or depend on confusingly similar string process. So just a suggestion on that.

SARMAD HUSSAIN: Thank you, Edmon. So we'll take the comments online.

UNIDENTIFIED FEMALE: [Yoshiro Yonaya]. "JGP status. JGP gathered seven people for initial member. JGP begun discussion from the end of the August and held three physical meetings in conjunction with e-mail information exchange. JGP is preparing formulation application statement to ICANN between JGP members, we are discussing whether Japanese has variants or not. It is not decided yet. This is because Japanese will be affected Chinese variants in TLD root zone. In Japan there is no authoritative variant list, so we are investigating several outcomes from Japanese variant definition projects. JGP is also investigating CGP rule draft.

"I have two questions to IP. Does allocatable strings be delegated to applicant by free of charge? And second question, as you know, Japanese uses multi-script such as Hiragana, Katakana, and Kanji and

also alphabet is commonly used, mixed with those scripts. Is it possible to include alphabet to Japanese script?"

ASMUS FREYTAG: So this was addressed to the Integration Panelists. The question of whether there is a charge for dual registration or not is outside the scope of this process. In case of the script mixing, a mixture of Hiragana, Katakana and Kanji is possible for labels under the Japanese script tag. The mixing of Latin letters would be excluded.

SARMAD HUSSAIN: We have one more comment online.

UNIDENTIFIED FEMALE: [Shakeil Amed]. "The EGIDS scale model has been criticized due to static and overlapping key questions approach to rate different languages, then how the EGIDS rating is being considered for to exclude langauges from IDNs."

ASMUS FRETAG: Response from Integration Panelist. In terms of ranking, we're really interested – we're trying to identify whether certain code points are used in a stable and widely-used and commonly and frequently used orthographies. There is no ranking system that ranks writing systems.

The EGIDS is an imperfect tool that we know is imperfect, but it allows us to have a proxy by giving us a quick indication of whether the underlying language is stable, is developing, is about to go extinct. And

if [inaudible] one of the languages that is not very stable, then we conclude from that that presumably the writing system based on it is also not stable.

However, there are many languages which are quite widely and vigorously used that do not have a standard orthography or not fully standardized orthography that we would have to potentially also exclude finally from the root zone until they are stable.

So the MSR that we have published represents only a first cut, and we deliberately cut it a little bit wide hoping to allow the generation panels to do the correct fine-tuning.

SARMAD HUSSAIN: Thank you, Asmus. And thank you, everybody, for joining us for the session. I hope it was useful. Please send us feedback on if you have any questions or queries unaddressed. Please feel free to e-mail us and we'll get back to you with more details offline. Please do send your feedback about the sessions in the morning and this one to us as well, so that we can further improve these sessions in the next ICANN meetings.

So thank you very much for your attendance and let's close the session. Thank you.

[END OF TRANSCRIPTION]