LONDON – IDN Variant TLDs Program
Wednesday, June 25, 2014 – 08:30 to 10:00
ICANN – London, England

PRESENTER: Good morning. We'll be starting with the IDN Variant TLD Session in about two minutes. Good morning. We'll be starting the session on the IDN Variant TLD Program Update. We apologize for starting a bit late, due to some technical challenges here. The way the Program is organized today is that we're actually going to present a Program update initially, on what the progress of the Program has been so far.

Then we have Integration Panel Members, who will present some details about the maximal starting repertoire, which was recently released. Then we have representatives from three community members, who will be presenting their current status. We have representatives from the Arabic Panel, Chinese, Japanese, Korean coordination effort, and a representative from Neo-Brahmi script. Then hopefully we'll have some time for questions and answers.

Let's start with the Program update. Just to give you a background, the IDN Variant TLD Program started back in 2011, when there was a significant need identified by the community. Before one could start thinking about what the solutions for IDN variants would be, it was necessary to first of all find out what the challenges were as far as variants were concerned.

The definition of variant changes across multiple scripts, so as the first phase what was done was a series of case studies, done through community effort, on six scripts – Arabic, Chinese, Cyrillic, [unclear

12:05] Greek and Latin – in which the communities went and analyzed their scripts, the challenges with their scripts, the kind of variants they may have, and produced six different reports.

Those reports were then – again, through a community effort – led by consultants. They were integrated into a single integrated issues report, which was eventually released in 2012. Based on that effort, the community deliberated on what the next steps would be and what was eventually agreed was to develop a label generation rule set for root, which will determine which code points would be allowed in the root zone.

In addition to that also, how those code points may create variants in certain scripts. A complete detailed procedure was defined to do that, and based on that, a procedure was created to develop and maintain the label generation rules for root zone, in respect of IDN labels. Work started on implementation on that procedure. At this time, that is with [Rea 13:18], so we are in the process called Project 2.2. We're actually implementing the LGR procedure.

Just to give you a bit more detail about what the LGR development procedure is, what is eventually needed? We need a single LGR for the root zone, however that LGR will contain characters or character sets from multiple scripts. To manage that, what was proposed was that there will be a Generation Panel for each script. The Generation Panels will be community-led efforts. They'll include Unicode experts, script experts, linguists, policy and other stakeholders from each of the script communities.

They'll come together and propose which characters in that script should go into the LGR in the root, and how the variants are going to be defined for those characters. Each community will propose their own part of the LGR. That will be the LGR proposal coming from that script community. All such proposals will be integrated into a central LGR by an independent Integration Panel. The Integration Panel will consist of experts from DNS, IDNA and Unicode, and people with linguistic expertize.

The Integration Panel takes on Generation Panel inputs and integrates that into LGR for the root zone. Therefore, it's important to realize that the Integration Panel actually reacts to the proposals the Generation Panels submit. The Integration Panel cannot develop the LGR by themselves. Just an update on what the progress is, as far as the project is concerned. We have quite a few communities that are now forming themselves into Generation Panels.

The Arabic Generation Panel was seated early this year, in February. The Chinese Generation Panel is in formation now. They have their proposals submitted and should be functional soon. The Japanese, Korean and Neo-Brahmi Generation Panels are being organized at this time, and they're also in the process of being seated very soon.

There's also individual interest from some other script communities, and that work needs to continue. These script communities need to have enough volunteers that they can organize themselves into a Generation Panel. As far as the work of the Integration Panel is concerned, as a first step towards LGR, the procedure states that a maximal starting repertoire needs to be released. That will be the input for the

Generation Panels to eventually propose a subset of the MSR for the LGR.

That work was completed and released earlier this year for public comments. The public comments were received, incorporated, their MSR was updated in that context, and it was released on 20th of June, right before the London Meeting. The MSR is out, and we have Integration Panel Members here who will go into detail of the MSR as well.

At this time, ICANN is reaching out to all of the communities that are forming themselves, or showing interest and assisting in the process of Generation Panel formation. Before going into more detail of the MSR, we also wanted to share some of the comments that were received. Basically, the community supported the conservative approach taken for the development of MSR.

There were a couple of queries about the process. Community identified need for addition in MSR in LGR script, for scripts that are already included, because in certain cases there are code points that script communities are not sure of. So they didn't want to include them at this time, but wanted to find out whether once they have the evidence of their use, whether they can include them at a later stage or not, if they give… [audio cuts out 19:00]

[TAPE CHANGE IDN-VARIANT-TLD-2-25JUN]

PRESENTER: …Seven scripts. There were queries about when those would be included in the MSR as well, for relevant GPs to form and do the work. There was also a question about MSR including some script that were not… There were no comments on those scripts in the public comment process, so it was not clear whether those repertoires have been reviewed by the community, and therefore whether it would be possible for communities to review it later on.

Then there was also comment about the need to reach out to additional communities so that people can get more involved, and the efforts that were going on in that context. There were a couple of content queries as well, on inclusion of specific code points, and concerning inclusion of languages spoken by small population communities. I will defer the response of the content to the next presentation, by the Integration Panel.

Very quickly though, as far as the process is concerned, to address the concerns raised by the community, what's being considered at this time is cyclical releases of both MSR and LGR, so communities in ICANN can organize their work. In that context, what would happen is that communities that have not been able to give feedback at this time, or have not considered certain code points, they can certainly do that in the next releases of MSR and LGR.

This process is going to be [additive 00:02:27], so even if you have reviewed a script and then you then want to add more code points for some valid reasons, and that addition does not cause any security and stability issue in the system, they can be added at a later time as well. There is an outreach effort going on by ICANN in this context.

ICANN has taken various measures to reach out to the communities, to help assist them in formation of Generation Panels. They're also trying to reach out to those communities that have not organized themselves yet, and trying to outreach to them to help them organize. At this time, MSR 1 is out. We're intending to try to cover many of the scripts, which are left out still, and have another version of MSR out towards the first quarter of next year.

The LGR work will start as soon as we receive proposals from the Generation Panels. Based on community input, if we receive Generation Panels early next year, we should have the first LGR out by the first quarter of 2015. Generation Panels will form later on, if they're not able to submit their proposals by early 2015, they'll then be considered in the future releases of LGR. That is at least what the plan is at this time, and we do request community feedback in this context.

We really do request the community to please organize themselves into Generation Panels and start this process. We still have many scripts to cover. It's just the beginning at this time. if you'd like to volunteer or help in the process, we'd be very glad to get in touch with you and get your input in the process. We also have a very detailed workshop on different panels that are organizing themselves and the challenges they're facing.

Some of that will be discussed in this session, but we have a more detailed workshop later on today. Please do find time and attend that. These are some of the references that will be available to you through these website. I'll stop here and hand it over to our other colleagues here, Michel and Nicholas, who are Members of the Integration Panel.

They'll work us through the details of the MSR, which was very recently released. Onto Michel and Nicholas.

MICHEL SUIGNARD: Good morning. My name is Michel Suignard. I'm a Member of the Integration Panel. I guess I'm there as a Unicode expert. I'm Unicode Secretary in fact, so I can represent Unicode interest or expertize in the Integration Panel. I'm also involved in the ISO 10606, which is the ISO side of Unicode. There is in fact some work that's done, especially in East Asia, that is synchronized with Unicode, so it's important to have a connection to the ISO side as well.

Also, in my former life, I did implement IDN for Microsoft, so I have a mix of Unicode and IDN expertize on the Integration Panel. Anyway, we've talked about the MSR at this point. MSR 1 was made available on June 20th, shortly before this meeting. You can go to the link to see the announcement on the various documents that do constitute MSR 1. That means work can now proceed for 22 scripts. That's the content of MSR 1. We'll go into more detail of what that means, to have 22 scripts on MSR 1, on the next slide.

The Generation Panel will receive from that a big repertoire from within the MSR. The MSR is… You can see that as [unclear 00:07:36] the characters are that you can use to add into the LGR. Obviously you have to do the variant analysis. That does not apply to every script, obviously. It's more concerned with CJK for the Han characters, and also Arabic may have these kind of issues, but many other scripts can pretty much use the code points and not bother with variants.

The ones with variants will have to decide if they want the variants to be allocatable or blocked. The decision to allocate is beyond the scope of us. We don't have any decision on that, but basically we're saying the Generation Panel will have to make that choice. We know that allocatables have their own set of issues, but that's beyond our scope. We will then generate an LGR proposal.

The Generation Panel will create the LGRs. They'll be submitted to public review and will then be integrated by the Integration Panel. Okay, so in numbers. There are 22 scripts. There's the list here. The main point that those are scripts doesn't mean that you may have more writing systems. There's some distinction between scripts and writing systems. For example, you may see Han between [unclear 00:09:26].

Han is obviously not just used in Chinese. It's used in Japan as well, and it may be used in Korea, depending if Hanja is included in the root. Then many of the list is used in India and South Asia. That includes the vast majority of scripts that are in use today, for in modern use. In fact, why we made that list, it was based on the gTLDs that were already being proposed. It's basically showing all the scripts for which there was interest today for gTLDs.

Obviously we expanded it a bit by adding related scripts. For example, you'll see that some of the scripts in India are include, that are not in fact where we have no gTLD, but because Devenagari was added, we basically added the rest of the Indian scripts as well. The same thing, between Thai and Lao. Thai has some requests for adding gTLD, so we added Lao as well, because Thai and Lao are related scripts. That's how we determined that list of 22.

Commonly [unclear 00:10:49] are just special script values that have to do with code points that are needed by multiple scripts. The total in numbers, we have 32,790 code points. That's a large number. In fact, there's even a larger number of PVALID. If you look at IDNA 2008 you can have almost 100,000 characters in IDNA. We reduced that somehow by having the Hangul syllable. You can see 11,000 plus, and then about 20,000 [unclear 00:11:26].

If you do the math, you'll see that in fact we don't need that much for the rest. I guess you could say that's a tribute to the efficiency of a lot of writing systems to be expressed in very few code points. In fact, the vast majority of reduction from 97,000 to 20,000 is in fact that we did significantly reduce the CJK set from everything that is PVALID to a set that is in fact today in use by TLDs for CJK.

NICHOLAS OSTLER: I'm the symbolic representative of the public commenting on this. The MSR 1 was submitted and made available for public comment, and the comment period was actually extended at least once. It ended towards the end of May. There are two sorts of comments we receive really, well, really three. One of two comments were just to say, "This is a very good job. Thanks very much. It's just what we required."

However, more comments directed to possibly changing the MSR were of two sorts. One was to say that we had omitted some code points that they would have liked to have seen included. In many cases, a quick look at the suggestions that we made showed that the suggestions were not PVALID, for example, or were in some way addressed to non-stable areas of the field of code points.

This was particularly the case of course in the vast field of Han characters. Nevertheless, we were asked in this revision in code point analysis to introduce some new characters. Certainly showing the flexibility of our judgment, we were able to include seven additional code points in the thing, but that is a very small number, compared with the tens of thousands, which you've just seen are included in MSR 1.

Another point made was that we were possibly in danger of possibly endangering more languages, by not including the scripts of minority groups. Now, we have actually addressed this issue, hopefully in an objective way, which is also open for revision, should the situation change. What we've done is to base our judgment on an independent scale, which comes out of the SIL – the Summer Institute of Linguistics – and has been developed particularly by Joshua Fishman.

If you are involved in sociolinguistics at all, you will certainly know that name. This is the so-called EGIDS scale – the expanded graded intergenerational disruption scale. It's intergenerational because it's concerned about the prospects of languages being propagated from one generation to the next. It has grades in it that show how endangered or how secure a language is, in terms of its prospects for having its intergenerational transmission disrupted.

It's a second version of the scale, which is why it's called the expanded graded intergenerational disruption scale. It's important to note that the primary consideration here is not population size, but what we could call the established vitality of the languages involved. Effectively what we're doing here is trying to find a way of characterizing the effective

demand for a writing system, which is not something that's widely done in the intellectual world, but it is the problem as it confronts us.

We have some probabilities of course, because some languages, which are the fundamental entities here, are written in more than one script, and sometimes the way in which they're written in those scripts, use some of the characters in unpredictable ways. Nevertheless, there is documentation of all these things. Written languages are above all things written, and so we can actually get evidence on these things.

What we decided was that in this intergenerational scale there are about nine major downgrades, you could say, that the best point at which we should characterize everyday use – the languages and the scripts that we're concerned about is that they should be in everyday use. That is what we inherit as our criterion.

From the procedure it seemed that that cut-off point came most conveniently between grade four, which is so-called educational use, where a language is in vigorous use, with standardization and literature being sustained through a widespread system, an institutionally supported education. It does seem to be a written language, which is well established.

As against a developing language, which is not quite at the educational level yet. It has a language that's used every day, certainly, and it does have a written literature. The literature is not propagated throughout the community though, the writing system is not used throughout the community, and also in many cases the precise use of characters has not been nailed down.

That's of course what we need for something to figure in the root. That's why we thought that was the right thing. We think it's objective. It is a system that's upgraded every three or four years by the SIL. Consequently it's possible for the status of a language to be revised, and its relationship to a writing system also to be revised.

We think that we had an objectively determined and independent scale here, which is nevertheless not set in concrete forever. It can be revised, but it does enable us to have a fixed target for any addition of the MSR. I think we can go to the next slide. Of course, I've talked about additions of the MSR, and this first MSR is, in principle, just what it says it is – MSR 1.

There are some other scripts, which are in widespread use, but for one reason or another were not included in MSR 1. Sometimes it was because the script, although small and stable, and hence eminently includable, did not seem to be represented by people who were concerned to be involved in the root at this stage. An example of that, in the early stage, was the Armenian script.

Others, such as some of the scripts of Southeast Asia, such as [Yanmar 00:19:56] and Khmer have had the problem that the digital encoding of the script is still disputed or unstable in various ways, and consequently we felt it would create unnecessary difficulties for the first stage to include them. Nevertheless, we are thinking of including them in the next edition of the MSR, which is likely to be related additively to MSR 1.

It's likely there will be more things. Nothing will be withdrawn. What will be included will be probably some more scripts, and also possibly some more characters within the scripts that we've already treated. We

are working on a timeline, which you can see here, and it is perhaps helpful at this stage to talk about how the different LGR proposals, as they come in, will be related to the MSRs as they are determined.

What this diagram is intended to show is that people could be working in a Generation Panel, and they could have started – and indeed they did start, particularly the CJK and the Arabic groups – before MSR 1 came out. It's also possible for groups to start now, now that MSR 1 has been determined. We would hope that a number of scripts will have proposals, which we very much hope will be in a position to validate, and which would mark this stage of LGR 1.

Some things might be started, which would not be ready to be validated in LGR 1. They would carry on, and indeed they might overlap into the period where MSR 2 has been [issued 00:22:16]. Indeed, if they are not validated in the LGR 1 phase, they will indeed need to wait for the validation that comes at LGR 2.

As we tried to show in this diagram, these Generation Panels could be started at any time, and there's sufficient flexibility built into the scheme that it's an advantage to get going as soon as possible, but you are not going to be cut off. I think that's the basic point from that diagram. Would you like to carry on?

MICHEL SUIGNARD: [00:22:59] The Integration Panel is obviously available as a resource to GPs, to make sure that their proposal is going to be successful. It's common interest to have the LGR to be accepted when they're being proposed, because if you have a rejection, that delays the whole

process. We oversee during the work of the Generation Panel, to help them, to make sure that the content is going to be at least within the guidelines of what we consider acceptable.

There's no guarantee received by us when we do that work. There is no black and white thing to say if it's going to be acceptable or not, but for sure, [unclear 00:23:41] we need to help with our expertize to make the LGR successful. One important point is that when scripts are related on that, it's very important that coordination between the GPs is done. That, for example, is the case for the CJK Panels.

There is this concern about variants, that there is need for coordination between the different communities working on the writing system based on the Han script. It's also true for things like Greek, Cyrillic or Latin, where we see what we call a homoglyph, totally identical characters, between Greek, Cyrillic and Latin. There's a need to have some coordination in that aspect.

We have a mailing list that is a way for us to be reached. We monitor that list and will answer any issues, questions or requests for clarification on the process on that list. It's archived and public. A quick think about the XML role on this process. The LGR expressed in XML may look at bit intimidating, but there's a good reason for that. First, it's shared with the work that IANA is doing in reformulating the specialized IDN gTLDs in that way, so we're in fact sharing the same tool.

It's also a way to be consistent and to have an easy way to have tools that do work on those XML representations, to make sure that due process is done when people want to check a domain or label, to verify that the given string is within… Okay. I won't go… I'm a bit out of time,

because we started late, but I'll see if you want more explanation. In fact, we have the expert on XML here in the room.

PRESENTER: Thank you Michel and Nicholas. We'll go right into the next presentation. That was work done by the Integration Panel, and work that's continued to be done by the Integration Panel will now move to community-based work, which is being done. It's just to give you an overview of some of the work here, we'll go into more detail in presentations by the community in the workshop session we have later on today in this room.

We'll start with the Arabic Generation Panel. It's being presented remotely by [Ahmer Masoud 00:26:42], who's calling in from Pakistan at this time. Over to him to speak about an update on the Arabic Generation Panel.

SPEAKER: Can you hear me? …Dr. [Sarmad 00:27:24] on behalf of the TFIDN. After his [journey] at ICANN, the Member of the [unclear] filled a big gap, which needs to be filled in the near future. I'm just presenting some of the words. Next slide please. TFIDN is a community-driven Task Force, which is working under the umbrella of the Middle East Strategy Working Group.

It was created with the objective to work on Arabic script LGR for a root zone – Arabic script internationalized registration data, universal acceptability of Arabic script IDN, technical challenges in the registration of Arabic script IDNS, operational software for the registry and registrar

operations, DNS security issues, with regards to Arabic script IDNs, and relevant technical training for the community, which is being done by the TFIDN all [unclear 00:28:43].

It was started in Kuwait, and communities are working further to work on it. The TFIDN currently has 36 members from 15 countries – not only from the defined region of the Middle East, but from other countries like Australia, England, Ethiopia, Germany and Malaysia. These members are speaking nine languages, like Arabic, Malay, Saraiki, Sindi, Pashtu, Persian, Punjabi and Torwali.

These members are covering many more languages. Also, the members are covering languages from East Asia, South Asia, Middle East, North Africa and Africa. These members have a diverse background, like academia, registries and registrars, national and regional policy bodies, and community-based organizations.

Membership of the Task Force is still open for all those who can contribute towards the community, especially if they have expertize in Arabic script and the DNS. Information regarding the Task Force is being updated regularly at the Middle East Working Group website, the ICANN community and Wiki space and on the ICANN main website.

In addition to that, email archives are also available at the Middle East Working Group website. Thank you to Fahd for efficiently updating the information on all these pages. Next slide please. These are the Arabic script TLDs that have been assigned or delegated up until today. I will read it, for the convenience of those who cannot read the script:

1) Algeria 2) Oman. 3) Iran. 4) The UAE. 5) Bazaar 6) Pakistan. 7) Jordan. 8) [Barat? 00:31:36]. 9) Morocco. 10) Saudi Arabia. 11) Sudan. 12) Malaysia. 13) Network. 14) Syria. 15) Tunisia. 16) Egypt. 17) Country. 18) Palestine, and 19) Site.

Next slide please. TFIDN has thoroughly discussed each and every code point one-by-one with full deliberation. It has discussed a total of 339 characters, even those that were actually disallowed by MSR and IDN 2008, to cover all characters. After discussed and deliberation, TFIDN has recommended 172 characters to MSR, for inclusion. At the TFIDN we are doing further working to reduce these characters for the LGR. So work is still under process, but it will be finalized within days, not months.

In addition to code points, TFIDN is deliberating on variants; keeping in view that it's important for security and stability and the possible DNS and phishing threats too. We have defined rules for inclusion, exclusion and deferral, for the code point and for the IDN variants. Until now, 120 plus cases of visually same or similar Arabic script characters have been identified by the group. It's been decided by the group that variants must not be allocated independently.

The security threats during the registration process are also being discussed within the group. The group is also discussion on usability in applications like browser, email, and other related issues. Talking about the progress, TFIDN has finalized Arabic script for a Generation Panel, and response from the MSR has been received. Principles for code point addition, deletion and deferral have been laid down, as I mentioned in

the previous slide, and principles for the variants, for the addition, deletion and deferral have also been finalized.

The different outreach activities are being performed, like At-Large at the Arabic IGF Meeting in Algiers. There was a presentation during the IGF in Bali. There was outreach during the Middle East DNS Forum. There was a presentation to the community at ICANN Singapore. There was a presentation at the APTLD Meeting. Now there's a presentation at the ICANN 50 London Meeting.

These are deadlines defined by the group, which have been decided on mutually to meet. It was decided during the face-to-face meeting, mailing list, and voice calls. We are expecting to finalize this LGR and variant rule by the end of this year. For us, for the Task Force, the deadline is December this year. That's when the group will hand over all the documents to ICANN, for public comment. Thank you very much. This is all from me.

PRESENTER: Thank you [Ahmed]. We'll keep moving on. Next we have a report on CJK coordination. I'll request Kenny Huang to present.

KENNY HUANG: Thank you. Good morning ladies and gentlemen. It's my honor to give a report on the CJK coordination. My name's Kenny Huang, Member of CGP. First of all, we have realized the problem of CJK is very complicated, and not to mention putting a CJK label into a zone, as that's even more complicated. Actually, it's not a question. It's a reality.

When you deal with that kind of situation, how can we put a CJK into the root zone?

First step to solve the problem is to institutionalize the structure, and that's the structure proposed by ICANN and the Integration Panel. That means in the Integration Panel, officially, we have C Generation Panel, J Generation Panel and K Generation Panel. All kinds of coordination should be under each individual Generation Panel, for example like a set-up CJK Coordination Committee, or a CJK Coordination Task Force.

That's what the proposal has been given by the Integration Panel. That's the potential structure we deal with, and although the CJK Generation Panel is not officially approved so far, we are heading this way. As realized, the problem we deal with – and it doesn't mean we can design the system without any boundary control – [unclear 00:38:48] we're going to have for CJK LGR.

For example, each CJK Panel creates an LGR, and an LGR including the repertoire and variant, including defined label permission, defined variant label, and [SI 00:39:00] disposition, such as allocatable or blocked. Given that we still have a tremendous coordination effort among the CJK community… If a LGR includes a Han character, the variant mapping must agree for all the Panels. The variant type may be different. The repertoire may be different. So that's a boundary condition we were given from ICANN.

Sorry. I was trying to speed up because we don't have too much time. I'll try to speak slowly. A lot of people came here to ask if we had any case to demonstrate what kind of problem we're going to have. I'll give them one very simple case, that overlaps between the Japanese and

Chinese communities, and probably in Korea as well. For example, like a Chinese LGR, the code point is one. Also, this code point is [six 00:40:05] in Japanese usage.

We have three sets of variant under this code point. We have variants one, two and three. They have different dispositions, for example like the first variant defined in the RFC 3743, last year it was allocatable. We called it [prefer 00:40:23] variant. The [unclear] of the two variants are just general variants. The situation we based on our operation was to block these two variants. Now we switch back to how the Japanese are running their variant.

Japanese one doesn't have any variant at all. Using the same code point – but we have different variant and different disposition between the Chinese and Japanese community – if we propose that kind of variant and rule set to the Integration Panel, like we're passing through the switch where [unclear 00:41:00] proposal or didn't integrate, if that kind of proposal didn't integrate it would be rejected and go back to the Generation Panel.

That's a problem we foresee and we're going to solve. I understand we have potential conflict among the CJK community, and initially we should have a very high-level conflict strategy. We didn't deal with every single code point, but we just have a rough idea of how we solve that kind of conflict situation. That could be many combinations. For example, in the original code-pointers we called "X" [six 00:41:44] in the CJK repertoire.

During CJK repertoire they have different [rows], for example like in Chinese we have [row 00:41:48] Chinese C. For Japanese we have [row]

J. For Korean we have [row] K. We have different rule sets. How do we integrate the three kinds of [rule] under register code point X? The combination could be that we could [accept 00:42:04] the original code point, and also abandon all the [rules] among CJK. That's one situation.

We can also adopt the X code point, and choose the intersection among CJK [rules]. We can also adapt X, choosing the union of CJK [rules]. Or, we could abandon X and also abandon all the CJK [rules]. Or, we could choose a [rule] based on a frequency of use. There are many kinds of pros and cons based on the conflict strategy we're going to apply. I'll mention again, that that's a very high-level conflict strategy. It didn't apply to every individual code point.

Okay, just reuse the same [rule], and choose what kind of strategy we're going to adopt. For example, like the code point one, which I mentioned in a previous slide. For Chinese we have three sets of variants. For Japanese, Japanese doesn't have any variants. If we choose union, and integrate label generation rules it will generate three sets of variant. If we choose intersection, that means all the variants will be abandoned.

So either way, either Han Chinese usage or Han Japanese usage, it doesn't sound very fair for both communities. Although, we have many years' IDN operational experience to [unclear 00:43:34] respect the ccTLD, but when we try to apply that kind of rule to the root zone, the situation would be totally different. Okay, what sort of methodology could we potentially use?

We tried to solve a problem. I didn't say it was the wrong way, just a potential way to divide and conquer. We tried to divide all the components into a very small sub-component, and solve each small

component by respecting methodologies. This is based on imagination, because we haven't started real coordination yet, although we don't have a CJK official Generation Panel established, so that's the potential way we can solve that kind of conflict.

Initially we can split [unclear 00:44:21] code point from the repertoire. Again, we don't have repertoire so far, so I choose the overlay code point from the IANA IDN repository. Under the IDN IANA repository, and the overlay between Chinese and Japanese, we have 6,181 code points that overlap. I didn't find any overlap among Chinese and Korean, or either among Japanese and Korean.

My assumption is that unless the Korean Generation Panel has proposed new repertoire, otherwise, based on the IANA IDN repository, there's no overlay code point among two communities. Based on that kind of assumption, which can refocus on the overlap between Japanese and Chinese, which is 6,181 code points. We need to define a respective conflict strategy.

From that [unclear 00:45:15], actually the problem will already be reduced. Probably more than 20,000 code points would be reduced to 6,000 code points. Okay, the next [issue 00:45:28] we're trying to do is we can do some kind of engineering, computation or design. For example, we can capture a source from the news – for example, like the Traditional Chinese [way 00:45:38], we try to collect the data from Apple News – and for simplified Chinese code points, we tried to collect the data from [unclear 00:45:45] news.

For the Japanese code points we collected data from [unclear] news. We tried to define the overlap code point, based on usage and

frequency of use. That's the computational result from the first page. We can [unclear 00:46:04] news code point from the overlap, and from the data we collect from different news – from Chinese Simplified, Traditional Chinese and Japanese – we found there are totally unused code points.

There are 2,730 code points that are totally unused. That means that kind of code point had already happened in our daily usage from the news. The overlap from the Chinese and Japanese usage is 1,312. The problem has already been reduced to 1,312 code points. Okay, finally, we need to compute in the frequency amongst these 1,312 code points.

That's the computational result – the top ten most popular words among [unclear 00:46:53], Traditional Chinese, Simplified Chinese and Japanese. Those are the top ten popular words. Lastly, the Chinese frequency of use. We chose [unclear 00:47:07] Chinese frequency of use is greater than Japanese frequency of use, just for demonstration purposes I chose the top 20, the top 20 code points you can see on the slide. The data says the total is 939 code points.

Also, we found a lot of Japanese code points, and actually their frequency of use is greater than the Chinese frequency of use, and the total code points generated is 363. That's the situation. The frequency of Japanese use is greater than Chinese frequency of use. We also find they exactly match. Chinese frequency of use equals Japanese frequency of use.

The generated data says there's ten. That means that the two data sets with frequency of use for Chinese and Japanese will be even.

That's the final outcome. Frequency of use is Japanese equals Chinese, and it's ten. The problem [domain 00:48:12] overlap becomes only ten code points that exactly match. After data processing and computation result, initially we probably have 90,000 – I call it 20,000 Han character code points – and we try to eliminate never-used code points. The overlap code point is 6,000, and trying to eliminate unused code points means there only remains 1,300.

After computation for the frequency of use, the overlap is only ten code points. The [problem domain] was effectively resolved, reduced. In the future, potentially, original overlap range can be redefined. For example, I only chose the exactly matched, and there's only ten code points that exactly match. If we have time and extra resources, we try to expand the standard deviation – for example, one standard deviation could be six days percentage that would be defined as overlapped.

That would require more intensive CJK coordination and deliberation. The main [unclear 00:49:21] 0.3446. Standard deviation is 0.55-something. Okay, not only do we deal from the computational side. We also deal with how we apply that kind of application. We can also use language tech when you submit a TLD application. For that kind of support, we can effectively reduce to validate a requirement from the market.

The source of language tech can re-reference a [unclear 00:49:53] 639. Okay. That's the end of my report. Thank you.

PRESENTER: Thank you very much. We will now move onto the last presentation for this session, by [Meha Gupta]. She'll be presenting the community work being organized for Neo-Brahmi scripts.

[MEHA GUPTA]: Thanks. I'm [Meha Gupter] and I'm a Member of the Neo-Brahmi Generation Panel. The Panel is not formally seated yet, so we're still in the formation phase. I'll begin my introduction with a little bit of an introduction of Brahmi and Neo-Brahmi, and then I'll give an update on what's happening on the Neo-Brahmi front. Brahmi is an Asian script that evolved during the final centuries of BC. It's an old script.

Most of the modern scripts in the Indian subcontinent have been derived from the Brahmi. For example, in India we have 22 official Indian languages. Out of 22, 21 belongs to Brahmi and one comes from the [Pacific 00:51:10] family. Geographically speaking, the script is being used in Central Asia, South Asia and Southeast Asia. Multiple language families use this script, largely by Indo-Aryan and Dravidian.

What is Neo-Brahmi? If we have Brahmi then why have we taken the approach of Neo-Brahmi? Most of the scripts are derived from Brahmi. Some of them are recognized. Some of them are non-recognized. Some of them are historic in use. We cannot cover all the scripts, so here we're only considering those scripts that are in modern usage.

Coming to the Neo-Brahmi Generation Panel composition, currently the group has ten Members. The Members have come from backgrounds such as linguistic, Unicode and academia. We still require more participation, more Members, to cover the diversity within the group.

We'll try to possibly cover all the major scripts and languages of the Brahmi family, but it's always subject to the availability of language experts and community experts, so we may drop some languages or scripts from the final proposal, if we don't get language experts for those scripts.

The group is currently working on gaining more participation within and outside India. Within India, we're approaching, one-on-one, two language experts, and getting the on-board. For outside participation we'll be holding a workshop in the upcoming APrIGF, which will be held in New Delhi, India, during August. We will be expecting much more participation after that workshop.

After the Singapore Meeting we put out our call for participation on the ICANN community website. Those of you who want to get involved in the Panel, you can send your replies to those email IDs. Though we're not seated yet, we have started our work. As a first activity we reviewed and commented on maximal starting repertoire. We had proposed some changes on the MSR and the Integration Panel agreed on some of the points and they included them in the MSR 1.

As I've already said, we're also planning a workshop in the APrIGF, to reach all the community for wider participation in the Panel. If we get the inner participation then we may submit the final proposal to ICANN by the end of August or early September. This is the Brahmi Generation Panel internal composition from a languages and scripts point of view.

Here you can see that Neo-Brahmi Integration Panel will have a Script Panel, and if the script contains more than one language then we may end up having a Languages Sub-Panel. We are seeing two challenges

right now. One is the inter-sub-panel communication. The Panel is too large, so it would be very difficult to communicate and coordinate between these Sub-Panels. Secondly, we'd require a large number of language expertize to cover all the languages and scripts.

As I've already said, if will don't get enough language expertize, then we may drop some of the languages or scripts from the final proposal. As [unclear 55:06] reason, we may have a reason in the future – like Neo-Brahmi GP 1, Neo-Brahmi GP 2, to cover all the languages and the scripts. This concludes my presentation. Thank you.

PRESENTER: Thank you very much. Here are some links, which are relevant for reaching out to the Generation Panels and to the Project Team at ICANN and also the Integration Panel. Now we can open the floor for questions. We still have about ten minutes.

JASON POLIS: Hi, I'm Jason Polis. I'm wondering about some of the variants that have been considered in the past, for example the Greek script. One of the first slides that was up was that Greek was considered years ago, but it's still showing up as yet to be done. Has it been done or hasn't it?

PRESENTER: There was an initial initiative where the issues around variants were identified for Greek script. That was one of the teams that worked on doing the first phase of the project. Again, as I said earlier, the first phase was determining what the possible issues are in each script, or

some of the scripts. Now the same community members need to come together to form a Generation Panel to propose a solution against those issues. This is a second phase of the same kind of work.

The issues for Greek were identified, but now the Generation Panel still needs to be formed, to propose the final solution, which will be integrated by the Integration Panel into the LGR. That step still needs to be done.

JEAN-JACQUES SUBRENAT: Good morning. This is Jean-Jacques Subrenat, a Member of the ALAC. During one of the ALAC meetings, Edmon Chung and Rinalia Abdul Rahim called for candidates to propose their services at various levels, especially for the Integration Panels. I did that. This morning I sent my candidacy, but hearing your very complete explanations I suddenly wonder whether there's any use for people who are neither computer technicians, nor really deep-down linguists.

I'm somewhere up there in the cloud with a vague knowledge. I have studied Chinese and Japanese and some other languages, but I'm not a computer scientist. Before I get engaged in this, is there any use for people lie me, who have dealt with languages in an international context, but neither as computer scientists, nor as analytical linguists? That was my first question.

My second question is, what do you think, in the Integration Panel, would be the average time allocation necessary for a volunteer? For what duration is it – three months? Six months? One year? That

question goes especially to Kenny, but I have also volunteered for Latin script. Thank you.

PRESENTER: Just to firstly clarify the terminology a little bit, the Integration Panel, the way we're referring to it, is the Central Integration Panel, not just the CJK Integration Panel. That's actually already formed and seated and it's already doing its work. As far as the Generation Panels are concerned, we do request input, not only from technical experts, linguists and Unicode experts, but also members of the community, in that process.

It's very critical and crucial to have community feedback into the process of Generation Panel formation, because at the end of the day, this work directly impacts how IDNs or IDN variants will eventually be used. That's a perspective that's very handy to bring on board, as far as Generation Panels are concerned.

Certainly, please do get involved, if you're even a community member and if you don't feel that you may have the requisite technical or linguistic expertize. I think just having an end user perspective on these Generation Panels is also useful.

SPEAKER: [01:00:46] I wanted to add the fact that the work process is subject to public review, so there are multiple stages where you can contribute as the public to review comments on anything that we do. Every stage, every submission, is going to be public and subject to reviews. Everyone in the public is totally free to make their own comments on whatever documents are proposed by either the Generation Panels or even the

[AGR 01:01:22] itself. There will be many, many opportunities for anyone to contribute to the process.

KENNY HUANG: Just two points. The first point, regarding the composition of each Generation Panel. Actually, either CJK Generation Panel [must be 01:01:36] requested regarding the member composition, and try to diversify the composition of the Panelists. Because we have the Panel Chair for the Chinese Generation Panel and the Korean Generation Panel, they are here, in [unclear 01:01:53] combination among all the Panelists.

The second point is regarding the implementation timeframe. Basically, we have propose a timeframe that, according to my memory, should be the 24th of October this year. Anyway, that's [along to 01:02:11] CGP. We need to integrate it with the JGP and KGP as well. For the potential deadline of having everything completed, I think it's very challenging, and we need to redefine the timeframe as well. Thank you.

PRESENTER: Any other questions? Okay. If you can find time, please do join us for the more detailed workshop on the work going on with the different communities. We'll have separate presentations on Chinese, Japanese and Korean, in addition to Arabic and Neo-Brahmi. The workshop will also allow for a more detailed discussion around some of these points that have been highlighted today.

Thank you very much for joining the presentations today. I'd like to thank all the Panelists here for making their presentations. Let's close the session. Thank you very much.

[END OF TRANSCRIPTION]