LONDON – IDN Root Zone LGR Generation Panels Workshop
Wednesday, June 25, 2014 – 13:00 to 15:00
ICANN – London, England

CYRUS NAMAZI: Thank you very much. We'd like to get started. Welcome to the session on IDN Roots on LGR Generation Panels Workshop. Here is the agenda for today's presentations. The aim of this workshop is to basically give a status of what is the progress as far as the Generation Panels formation and work is concerned. We would have presentations on the Generation Panels being formed by different script communities. We have on our agenda presentations from Chinese community, Japanese community, Korean community; some coordination efforts going between the Chinese, Japanese and Korean communities, and the status of Neo-Brahmi community efforts going on, and finally an update on the Arabic Generation Panel.

Basically, this project is called Project 2.2. It is focused on developing Label Generation Ruleset for those Root Zones. The Label Generation Ruleset will eventually be used to determine which labels can go into the Root Zone and also determining whether those labels will have any variance and whether those variances would be allocatable or blockable.

The process is basically, it's a two-step process. In the first step the communities organize themselves into what we call Generation Panels.

These Generation Panels organize around each script and a pool of experts and community members discuss and finalize a proposal for their own script community. Basically suggesting which core points, or

subset of core points, in their script should be allowed or used in the Root Zone for labels. That proposal goes to a central panel called Integration Panel, which takes proposals from each of the script communities and integrates that into a single Label Generation Ruleset. That single Label Generation Ruleset will eventually be used for label generation for the Root Zone.

We have here members of the communities which are actively developing or are part of different Generation Panels. We also have a couple of members from the Integration Panel. Today we will go through presentations for these scripts. After that we'll have a question and answer session in which you can direct your questions to either the Generation Panel members or the Integration Panel members. Let's get started.

We are still missing a couple of people here for presentation, so we will start with the Korean Generation Panel presentation and the CJK coordination report and go on from there. Once we have the other speakers we'll come back to the Chinese and Japanese Generation Panel. Let's move on.

We'll start with the work which is being undertaken by Korean Generation Panel.

KIM KYONGSOK: My name is Kim Kyongsok representing Korean Label Generation Panel. The KLGP met just once. We are in the stage of starting. I'll tell you the status of KLGP, and secondly, the result of the KLGP meeting. Then I

attended the CJK meeting in Shanghai and will tell you a little bit about this and other issues and tell you the future plan of KLGP.

The status of KLGP. Currently we have about ten members. There are linguistic experts, policy experts, community members and registry members. We are still trying to recruit members who have expertise in culture. As I said, KLGP met just once and I will tell you the result.

We discussed what will be the scope of KLGP regarding Hangul syllables since only modern 11K syllables are included. Since [then] there is no variant issues. The range is shown here: AC00D7A3. Regarding Hanja, well, there can be opinions, so we need to make consensus. Until now we did not have consensus yet, but after returning we will meet again and will try to make consensus about including Hanja.

We reviewed the MSR-1 and we did not make any comments during comment period. When I came to know the existence of MSR-1, the comment period already ended. Currently MSR-1 includes 11K Hangul syllables, which are suitable for representing modern Hangul syllables.

About Han script. It's set to include CJK unified domain, extension AP. Since IICORE is a subset of CJL Domain extension AMP, it does not show any extra characters.

I want to know how the Han characters in MSR-1 was created since it includes CJKU main extension AMP, but I'm not sure. That was a question we had during the meeting.

I attended the CJK meeting in Shanghai and got some information. It was held at the end of May in Shanghai. About 20 people attended. From Korea, I attended. Japan and USA attended by teleconference.

During CJK meeting in Shanghai, MSR-1 was reviewed and I asked about the status of Hanja domain in Korea. Currently in Korea we have .KR and .Hangul domains. Under these two domains we do not allow Chinese characters. We allow only Hangul syllables in addition to English letters, digits and a hyphen. KLGP wants to participate in the CJK meeting in the future too.

At the CJK meeting in Shanghai there was comment requesting to remove about 89 Idu characters from the Han list. After returning from the Shanghai meeting, I checked the characteristics of 89 Idu characters. Some Idu characters included in IICORE and it seems desirable to retain IICORE in the MSR-1. I suggest not to exclude Idu characters from MSR-1, just because they are labeled as Idu. Instead, we need to apply some other principles or criteria [inaudible] not to exclude some Idu characters.

For your information, the term Idu does not show up in ISO 10646 in third edition anymore. In the past there was such a term, but now it is gone.

Within KLGP we discussed other issues. What will be the scope of KLGP? Probably need more time to consider the scope. For example, are we only going to include Hangul in KLGP and participate in CJK LGP, or within Korean label are we going to include both the Hangul and Hanja characters and participate in CJK LGP continuously? We need to consider more.

I guess it was already answered. You could include both Hangul and Hanja characters if KLGP wanted to. It was prepared before coming to this meeting and this section need not be discussed anymore. I'll skip it.

In the case of Korean language, most populations reside in South Korea and North Korea. There are some other people speaking Korean, but mostly they follow South Korean or North Korean conventions. So we need focus on South and North Korean conventions.

Unfortunately, it is very hard to make North Koreans involved in KLGP. I don't know if there will be any good solution, but as an alternative we could probably make North Korean refugees residing in South Korea involved in KLGP. Of course, they do not represent North Korea, but currently, since there may not be a good solution, it might be one alternative. If you have any good suggestions, please let us know.

About future plan, after this London meeting, we will have a KLGP meeting and we plan to have a public hearing, then we want to draft KLGP proposal. Thank you.

CYRUS NAMAZI: Thank you very much. Our next presenter, can I request the Japanese presentation?

UNIDENTIFIED MALE: Okay, just one page because we haven't established any formation of the GP. This is, I got off schedule, which has been considered only by the Japanese candidate on the panel. Of course, we have to arrange the schedule among the CJK cooperation.

At this moment, we are now – it's June, and we are contacting two candidates on the panel. Who we are contacting is mainly those who were on the panel. When second level IDN .JP was introduced 13 years

ago, and 13 years ago the panel consisted of a lot of various experts from language experts and, of course, the engineers and so on. So they can be a core of the new Japanese Generation Panel, so we are contacting them.

We explained to them the difference between the second level IDN and the .JP and the LGR. I hope they understand the difference and I hope they understand the new issues when we consider the top level domain IDN.

We are still seeking candidates and hopefully in two to three weeks we can make a pre-meeting of the core team. After that we will submit a proposal of the candidates of the Generation Panel to ICANN. We will have a monthly meeting for six months. This is our expectation and the optimistic schedule. We want to do that.

After considering draft rule among only Japanese, we will talk with the CJK people how to coordinate among us. This is just all I can say at this moment. Thank you.

CYRUS NAMAZI: Thank you very much. I guess we'll keep moving ahead and talk about CJK coordination issues and work being done in that area. So we'll move onto the next presentation. My apologies for jumping around between the slides, but I guess we're just trying to sequence the presenters who are available. Let me move to the CJK slides.

KENNY HUANG: Thank you, again. This is Kenny Huang. I apologize for the repeating talk. What I say will be the same as I did this morning. As we know, CJK is a very complicated issue, especially when we put in CJK-A script into a Root Zone. When we come with our own kind of [variant] and our own kind of disposition, that will make a lot of difficulty in the Root Zone operation. So we need to find a feasible was for the Root Zone operator to see how CJK can be harmonized and put into the Root Zone.

This is the proposed structure for the integration panel. We have the individual CJK panel, and underlying, there will be a CJK Coordination Committee. So we are waiting for K-Generation Panel and J-Generation Panel to formally establish. Once they're established we can start to work on the coordination stuff.

There's a boundary condition for the CJK Label Generation Rule and each CJK Panel has to create a Label Generation Rule. Each Label Generation Rule includes a repertoire and variant. Especially we need to define the label permission, define the variant label and assign disposition, such as allocatable or block. We require tremendous coordination effort among CJK community, especially if LGR includes Han character.

A variant mapping must agree for all the panels. So far we have CJK Panels. Once we create a variant, that variant mapping must agree from CGP and JGP and KGP. The variant type may be different among the three Generation Panels. The repertoire may be different among the three Generation Panels.

[inaudible] easy to understand what could happen with the overlay case. For example Chinese Label Generation Rule for the character set in

terms of English stands for one. Japanese also using the same character set.

For this core point, Chinese has three sets of variants, but Japanese doesn't have any variant. Once this case, both proposal submit to the integration panel, but [inaudible] say what be different and disposition is different. That would not be considered an integrated proposal. It would be rejected and sent back to Generation Panel.

Based on the potential integration requirement, we can come up with some sort of combination conflict strategy. For example, we can consider X stands for overlap [inaudible], and RC stands for Chinese Generation Rule and RJ stands for Japanese Generation Rule. RK stands for Korean Generation Rule.

For example, I've got a first [idea] – we can adopt X code point and abandon our Generation Rule among CJK. Or we can adopt X original [code point] and we choose intersection among CJK Generation Rule. Or we can choose to adopt original [code point] X and choose, try to apply the union of CJK Generation Rule. In the worst case, we must abandon CJK Generation Rule. The last one could be adopt the rule based on the frequency of use.

Of course there are different kinds of pros and cons among all kinds of conflict strategy. I didn't say any of the strategies would be the best or would be a single way to solve the problem. We're just trying to deliberate all kinds of strategy we can use to solve the conflict.

Again, we go back to the same case. If we choose the union among Chinese Label Generation Rule and Japanese Label Generation Rule for

the code point 1. When we choose union, the original code point will remain and that will come with three variants based on the union proposition.

If we choose intersection between these two Generation Rules, finally that rule will apply, so there's no variant for the original code point. You can see from this example, it depends on what sort of strategy we use. If we choose union, that seems not very good for Japanese community. If we choose intersection strategy, that seems not very good for the Chinese community.

That's the way we try to solve the problem. Divide the problem into different subcomponents and choose a proper way to solve each subcomponent and find a proper solution.

For example, initially we can [inaudible] an overlapping code point from the repertoire. The repertoire I chose here was the IDN repository. From this repository we didn't find any overlap between Korean Han character and Japanese characters. But there's the overlap. The overlap code point is 6,181 between Chinese code point and Japanese Han code point. I exercised a kind of approach and the problem [domain] reduce to 6,181.

The other thing we're trying to do is collect the data from different kind of median. For example, we collect traditional Chinese data from Apple news, and we collect simplified Chinese code point from Sina News and we also choose Japanese data from Mainichi News. After collecting the data from various sources, we try to compute to get a usage and try to compute to get a frequency of use.

That's the first outcome from the initial computation. We split used code point from the overlap. From the diagram you can see Chinese character and Japanese character usage overlap only 1,312. There are 2,739 code points they didn't use at all from the data we collected. Probably when we expand the data from a more wider range, the simulation will have different results.

Second stage, we try to compute the frequency of use from different code point. I just tried to simplify to demonstrate what those results are going to be. Let's demonstrate top ten most popular words from traditional Chinese code point and simplified Chinese code point and Japanese code point.

This diagram shows the Chinese frequency of use, and it's greater than Japanese frequency of use. The data we calculate has generated a data set of 939. That means we have 939 code point frequency of use in Chinese is greater than the frequency of use in Japanese.

In reverse, we can also find out the frequency of Japanese use is greater than Chinese frequency of use. The data set is 369. We can also get even the frequency of use between Chinese and Japanese. The generated data set is 10. So this 10 code points, actually, are the same frequency of use between Chinese and Japanese. That's the final outcome. We get 10 code points that exactly match between the usage in Chinese and the usage in Japanese.

In the future we probably can redefine the [overlap] range. For example, we can, in this case, our use is exactly matched. In the future we can probably have some margin to expand, to really define the overlap to a certain level of standard deviation. Once we extend that

kind of definition, that also requires intensive CJK coordination and deliberation.

From this diagram, in the right hand side, it's [more close to the rule] apply for Chinese Generation Rule. For the standard distribution, on the left hand side was more close to Japanese Generation Rule. Also, we should reconsider the language tag when people try to submit a TOD registry. Thank you. That's the end of my presentation.

CYRUS NAMAZI: Thank you very much. We'll keep going forward. The next presentation is—

ANDREW SULLIVAN: [inaudible] on that particular topic.

CYRUS NAMAZI: We can certainly stop and take a question if…

ANDREW SULLIVAN: Okay. So my name is Andrew Sullivan. The thing that I'm struggling with this particular topic is that there seems to be an assumption that you need exactly one rule for every Han character. That's not the way the procedure says. You can have more than one rule because they're scoped by the language tag. Therefore, the problem that was just described isn't a problem. You don't have to have a single output for all Han characters. You can have Han character output in different linguistic contexts. We spent a lot of time on this thing during the

development of this procedure, so I don't understand why you're not using that feature of the procedure in order to make this problem go away. You don't need exactly one rule for every Han character. What you need is a rule for Han characters according to the language tag of the label at submission time so you can shift back and forth.

What that means is that, for instance, for a Japanese use of a Han character, it generates the variant, but they're blocked. For instance, a traditional character would then have a simplified variant in Chinese that would be activated if the label were a Japanese character instead and it had the same Han character. It would generate the same simplified Chinese character in its variant set, but that variant is then blocked instead of allocatable. That's the way this is supposed to work. There isn't a rule that everybody has to agree on exactly what the variant set is and what labels you get out of it. I've watched the presentation twice and it seems to me that there's a fundamental error in its use. It could make your life simpler.

CYRUS NAMAZI: Does anyone on the integration panel want to clarify this point?

UNIDENTIFIED MALE: I agree with Andrew. No controversy here. We agree 100% with Andrew.

UNIDENTIFIED MALE: My assumption is one script, one rule. Only one rule exists in the Root Zone. That's my understanding. If this understanding was incorrect, then we don't need too much coordination at all.

UNIDENTIFIED MALE: Again, you need coordination to set the variants, but how you deal with the variants is separate for each panel. The only thing you have to agree specifically to set the variants is that intersect which is a different repertoire than you are creating. You have to make sure that you look over the same set of variants do exist for one character as you are putting in your repertoire. Even the rules that you put with that variant can be different for each GP. They don't have to be the same.

For example, that was mentioned by Andrew, in one case it can be allocatable, or in one case it can be blocked. That's a [inaudible] rule. So, the roots are not shared among coordinated sets. Only, to some degree, the variant set is shared. In fact, even then they're not exactly the same. They may be a subset depending on the size of the different sets created.

UNIDENTIFIED MALE: Actually I think Kenny just proposed a solution about the coordination. I can give us a suggestion that we treat the variants, we should set up a unified repertoire, especially about the mapping relationship between variants. We can set different types for different characters. Maybe for a character that we think is variant in Chinese language community, but for the Japanese. Anyway, they don't regard this variant and they can just block it. It's not allocatable.

Because I think KGP and JGP has not been set up formally, and before we have a consensus about that principle, that is what [inaudible] give us, and Kenny just proposed another solution. What about two different solutions? An intersection and the union solution to solve this problem and we do some [statics] jobs about the frequency of the character used to decrease the complexity of the character set union? That is my understanding.

CYRUS NAMAZI: Any further comments from the panel? Nicholas?

[NICHOLAS OSTLER]: Another comment to make, in fact that was mostly for the Korean but we did not have a chance to give feedback after the Korean presentation.

I just wanted to comment briefly on the MSR-1 content for [inaudible] repertoire was done because there was a question apparently. The answer is that we did take the .asia Chinese set (.CH) and we also added the.JP, in fact easier for the Japanese TLD or the .asia and .JP are the same content so it does not matter. Then we union that with IR cross set that which is another collection of CJK characters that are both defined in both Unicode on 10606. On that added one character.

It's pretty well defined. We basically practice from the existing TLDs that are created in our repertoires today. We defined the sandbox. What we [see] doesn't mean that every Generation Panel will take everything because it's a union repertoire, so we would expect each of the Generation Panels to, in fact, subset back to probably what they had

originally. That's more-or-less their own decision what they do within that sandbox. [Inaudible] takes original repertoire, is it the .asia, .CN, or less or more. That's going to be their decision anyway. I just wanted to explain that this is how MSR-1 Han content was created.

That's probably the way of the document. The Integration Panel, we are basically part of the MSR deliverable where the document that described MSR-1 was created. That's part of it.

CYRUS NAMAZI: Thank you. Let's keep moving on. We'll come back to this discussion if there are more queries. On to Neo-Brahmi work on the progress. Neha Gupta is here to present that.

NEHA GUPTA: Thanks so much. First of all, my presentation will be a little bit repetitive for those of you who were here in the morning session. I'm Neha Gupta and I'm a member of the proposed Neo-Brahmi Generation Panel. Along with me are my friends from the Neo-Brahmi Panel are also here in this room.

Coming back to the presentation, I'll start with the direction of Brahmi and Neo-Brahmi. Brahmi is an ancient script which evolved during the final centuries BC. Brahmi is a modern name given to the oldest script which is being used in Indian subcontinents. In India, we have 22 official languages. Out of 22, 21 have been derived from Brahmi. Geographically speaking, the scripts are being used in Central Asia, South Asia and Southeast Asia.

Brahmi uses Abugida as a writing system where the writing system treats consonants and vowels as a separate unit. This diagram depicts how these modern scripts are derived from Brahmi. Here the [inaudible] is modern scripts, like Devanagari, Bengali and Gujarati you can see. Gupta, Grantha and Kadamba, they are all families which are derived from Brahmi.

Even though the shape is different from script-to-script, the basic philosophy is always common. They all are akshar driven. Akshar is a basic syllable unit which treats consonant and vowel sequence as one unit. So we represent those labels as a syntax in akshar. The syntax is very much important when representing the scripts in digital medium, even in the case of Internationalized Domain Names. Brahmi [inaudible] many scripts. Some of them are recognized, some of them are not recognized. So we are not covering all of the scripts. Here we are focusing only on the modern scripts which are in modern usage.

We do have a past experience. We have been working on IDN ccTLD. Some of the members are involved in IDN ccTLDs for India. So we have 22 official languages when we consider a single script in multiple languages. When we include multiple scripts and [one] language pair, the number becomes 27. Each language there is having its own code point set, variant and complex whole label evaluation rules. In the IDN ccTLDs, we have similar looking characters as a variant, but we are dropping them out from the LGR. So those will not be treated as a variant in LGR. We will try to simplify the whole label evaluation rule from complex to simple.

Current members list. Currently the group is of ten members. They have come from backgrounds like linguistic, Unicode [academia] and these are the members names.

This is the internal composition in terms of scripts and languages. In the MSR-1 we have 22 scripts. Out of 22, I think ten have come from the Indian scripts. Currently we are not focusing on all ten scripts, so we'll only be starting with six scripts, which either have ccTLDs or gTLDs. If we get language experts, then hopefully we will cover all.

The group is currently working on gaining more participation within and outside India. We are approaching two language experts, two [community] experts and getting them on board, but it is very difficult to explain the process, explain the documents and how it would benefit them. Participation is coming slowly, but hopefully we will get [inaudible] participation by August or September.

For participation, we have floated a call for participation on the ICANN community page, so please go to that and if you want to get involved, please keep in touch with us.

As a first activity, we reviewed and commented on Maximal Starting Repertoire for the Root. We had some proposed some changes in the MSR and IP Panel has agreed on those changes. For outreach, we are planning to have a workshop in AprIGF for reaching out to the community for the wider participation in the panel. Hopefully after this workshop we will get enough participation. Then we will be able to submit the Neo-Brahmi Generation Panel proposal to ICANN by end of August or early September. Thank you.

CYRUS NAMAZI: Thank you very much. We will now move on to Arabic Generation Panel. We have Ahmed Masood on remote participation. Ahmed, are you online? Can you hear us?

AHMED MASOOD: I am online, can you hear me?

CYRUS NAMAZI: Yes, we can. Over to you.

AHMED MASOOD: Okay, thank you. Thank you very much. I apologize for those were also part of the meeting in the morning because it is just a repetition of the information. I will briefly explain an overview and progress of our taskforce on Arabic IDN, [inaudible] TF-AIDN. As I said in the earlier meeting, someone would be presenting on behalf of the TF-IDN in different forums and I'm just a backup for him and for sure not [of that] quality. Next please.

TF-IDN is a community-driven taskforce which is working under the umbrella of Middle East Strategy Working Group. It was created with the objective to work on Arabic script, LGR for Root Zone, universal acceptability of Arabic IDN, technical challenges [around] registration of Arabic script IDN, operational software for registry and registrar operations, DNS security issues relevant to Arabic script IDN and relevant technical training for the community which are being arranged by the taskforce. Initially it was arranged at [inaudible] this year and TF-

IDN is planning, in fact it is in coordination with this strategy working group. It is planning somewhere else in the region. Next please.

TF-IDN has currently 26 members from 15 countries. Not only from the defined region of Middle East Strategy Working Group, but also from the countries which are not part of the region like Australia, England, Ethiopia, Germany, and Malaysia. These members speak nine languages – I'm just talking about the languages which relate to Arabic script, not the countries mentioned earlier.

These languages are Arabic, Malay, Saraiki, Pashto, Persian, Punjabi and Torwali. These members cover East Asia, South Asia, Middle East, North Africa and Africa. These members have come from different backgrounds like academia, registries, registrars, national and regional policy bodies and community-based organizations. Next please.

Membership on the taskforce is still open for all who can contribute, whether linguistically or whether technically. They have expertise in policy-making, or they have expertise in DNS working.

You can see the information appearing on your screen. It is being updated by Fahd very frequently. This is available on the Middle East Working Group website and the ICANN community website. All of these are very useful information. Next slide please.

These are the Arabic script TLDs, which have been assigned or delegated until today. I will just read it out for the convenience of those who cannot understand this script. (1) Algeria; (2) Amman; (3) Iran; (4) Al Bharat; (5) Bazaar; (6) Pakistan; (7) [Al Jordan]; (8) Bharat; (9) [inaudible]; (10) Al Saudia; (11) Sudan; (12) [Mlessa]; (13) [Shubukta];

(14) Syria; (15) Tunis; (16) [inaudible]; (17) Kurta; (18) Philistine; and finally (19) Morocco. Next slide please.

These are the outcome of our first meeting and part of the second face-to-face meeting for TF-IDN. There were total 339 connectors, codes in fact, which were discussed entirely by the group, including those which are initially disallowed by the MSR and IDN 2008. Just to cover all languages who have been represented in the group or are not in the group. Our members are covering more than one language for their own region.

172 codes are submitted to the MSR and then we are further working out to reduce these connectors for the security and stability to maybe 150 or maybe 125. What will be the result, it is still in the process and will be finalized within days. Next slide please.

In addition to previously discussed TF-IDN is also deliberating on variants because these are very important for the security and stability point of view. We are also considering the possible DNS and phishing attacks on the domain names because of the usability of these variants. The group has identified 120-plus cases of same or similar Arabic characters within their meetings and their mailing list or the conference calls. It has been decided that the variants must not be allocated independently.

Security threats during the registration process are also discussed very enthusiastically. Discussions may be [inaudible] of our group. It is maybe an offense to exclude some of the characters, but we are considering the security and stability as devised by LGR as the first priority.

The group is also discussing the usability of the applications, like browsers and e-mails and searching and privacy, because these will be the applications most frequently used after the introduction of Arabic IDN. Next please.

TF-IDN has finalized Arabic Script for Generation Panel and a response from MSR has been received to the TF-IDN. We have finalized the principle for code point addition, deletion and deferral cases.

Similarly, we have finalized principles for variants for addition, deletion or deferral. These have been finalized and we are also working for the outreach of the community. Different activities have been launched by the group. It was started at an IGF meeting in Algiers with a presentation during the IGF in Bali outreach during the Middle East DNS forum, presentation to community at ICANN Singapore and presentation to APTLD meeting and a presentation to this ICANN meeting. Thank you. Next slide please.

We have discussed in the group and finalized the deadline. We hope to finalize all of our LGR related work by the end of this year. We will be submitting to ICANN integration panel and it will be released for public comments. Thank you very much. This is all from me.

CYRUS NAMAZI: Thank you, Ahmed Bakhat Masood. Now we will move on to our last presentation on Chinese. That was one of the first ones on the schedule. You will have to excuse me; I have to jump back to the first slide.

WANG WEI: Sorry that we have to jump back to the updates from the Chinese Generation Panel. The first slide is about the evolution of Chinese characters. I don't want to mention it too much, but what I want to say is that the modern Chinese characters was formalized in 200 AD, and since then nothing changed to the script. It was used for the following thousand years and that's what we use now.

It was used widely in East Asia, what we call Asthenosphere. But now it was just used in several countries, including China, Macau, Taiwan, Hong Kong, Singapore and Malaysia. Of course, the Chinese character is also used in Japan and Korea, formally or informally.

The target script and scope. The target script, of course, is Hani, including Hans, which means Hans script simplified and Hans script traditional. Also in Japan and Korean script, there are some Chinese characters. The relationship between Han character and Japan script and Korea script is just like this graph.

It is fortunate that the complete character set for the Chinese script in IANA IDN table from .CN and .TW and .Japan all fall in the range of MSR. We noticed that in the IANA IDN table from .KR there is no Chinese character, so there is no intersection between us. That doesn't mean for the Root Zone, for the TLD there will not be the Chinese character. It depends on the output of KGP from the Korean community.

The best practice is by the CDNC. CDNC is an organization that opened in 1998. The founding members include two Japanese, a Macau and Hong Kong. In 2004, CDNC submitted to IANA a unified Chinese character set. The registration sent to IANA based on the RFC 3743 and 4713 and the CDNC put forth the guide rules for registration for the

second level domain, which has been applied in above TLD for the past 12 or 13 years. We call it the best practice. That's why we would like to propose a CDNC character set to be treated as an initial character set of CGP. That's a good foundation of our work. This initial character set, of course, all fall in the range of MSR.

We have gathered 13 experts from a variety of backgrounds and of seven different countries. Let's have a look.

The first one is Chris Dillon, from UK. Of course, he is a linguist and his study covers Chinese, Japanese and Korean and he can speak Japanese and Korean and some Chinese, of course. We have Chao Qi, the engineer from CNNIC. His focus is on the registration system. He's an engineer. Di Ma from ZDNS from China. ZDNS is a registrar of ICANN.

Professor Lee Guoying, the professor from the Beijing Normal University on ancient Chinese character. Jiagui Xie from CONAC, the registry of two Chinese New TLD registry, also in charge of the registration system. Professor Joe Zhang from the Beijing UniHan Digital Technology Corporation Limited. His study covers Chinese, Japanese and Korean. He is also one of the authors of the CDNC character sets.

Jonathan Shea from the HKIRC.hk. He is the CEO of .hk. Joseph Yee, from Afilias. I think his nationality is Canada. He can speak Chinese very well and he is familiar with Japanese. Kenny Huang, from TWNIC represents TWNIC and the Taiwan community. Linlin Zhou, also from China, from CNNIC. Linlin Zhou contributed to the integrated report of the variant study.

Nai-Wen Hsu from TWNIC, engineer. Ryan Tan from Singapore as a representative from SGNIC. Wei Wang, that's me, I'm the CEO of CNNIC. Xiaodong Lee, from CNNIC. Zheng Wang, from the CONAC. Zhiwei Yan from the CNNIC.

I actually discussed with IP with Han Chuan yesterday, we noticed that since Malaysia is an important area of Chinese language, we need to add one more expert who will represent the Malaysian community. So I will ask the Malaysian community to provide an expert.

The relationship with past work. Actually, for Chinese, for CGP, if there's no overlapping issue between K and J. It will be quite easy for us. Since 1998 we have started some work on the second-level domain and JET is a joint engineering team about the Chinese character for China, Japan, and Korea. In CDNC we focus on the traditional character and simplifying the character equivalence issue. Our solution has been applied to .CN, .TW, and [.HK] for the past decade. This solution was approved perfect.

The next slide is about our work plan. In this generation, we just started the CGP work. We think our goal is to finalize, finish all the jobs between the ICANN 51 in Los Angeles. But we found it's pretty hard to finish that aim because we are struggling with coordination between KG and C, especially about the overlapping Chinese character set. We provided the first revision our proposal in April. Now we have found there are many questions, many issues, we haven't considered. We got comments from ICANN and we just revised the proposal in the past months. We just submitted a second version yesterday.

In the past four months we held two joint meetings of CGP and CDNC. We also invited Professor Kim from Pusan University from Korea to join us in Shanghai in May.

I mentioned that it is very difficult for us and we are struggling with this job. I think there are two challenges. The first is about the character reduction or extension. The second is about coordination. I notice that ICANN suggested that we should keep the consistency between TLD and the second level. That means if it's a compulsory requirement, that means we'd better keep the character CDNC character set unchanged and don't modify it because that's the best way to keep our current operation – the registration under .CN (.China), .Taiwan (.TW) the same even after we create Label Generation Rules for Roots.

We got some suggestions from the linguistic experts that maybe it is necessary to modify the current CDNC table. For example, some absolute Korean Lidu character could be removed from the current CDNC table. For example, some Han radicals/strokes that are not regarded as independent Hanzi could be, or should be, removed from the current CDNC table.

These modifications will make the table more reasonable, or more rational. But we are worried that if we do that job, in the future, Chinese repertoire in the Root Zone will be different from the character set in .CN and .TW. I'm not sure if that will cause inconsistency problems for us. That's one of our concerns.

Another is about the coordination. Here we have some graphs that I got from the Internet, which means that there is a lot of Chinese characters used in Japan on the TV and in the newspaper.

I was told that there are over 2,000 Chinese characters formally or legally used in Japan. For Korea, I was told that the Chinese characters are not allowed in formal documents or in official documents. Thanks, Professor Kim. He provided some examples of the Chinese characters used in some cases, for example in the newspaper, the restaurant's brand name, or in some books.

So thanks for the experts who provide a solution about the coordination in the Shanghai meeting, which means for the CGK panel, in their repertoires, if a CGP includes a Han character, the variant mappings must agree for all [three] panels, but the variant types may be different. If CJ and K both agree with that principle, that would make it easy, I think.

So far I think we can say that the Chinese Generation Panel are okay with this principle. We just want, with the help of ICANN, to make a consensus together with K and J.

Kenny Huang just gave us another solution. That is because before we get this suggestion from ICANN we think maybe there should be more solutions to solve the coordination issue. That's all, thanks.

I also want to say that our work plans are not changing yet. We want to finish our other job between the next ICANN meeting. That really needs the help from ICANN and from the K and J. If we can have an agreement on the repertoire on the variants and the coordination principle, that will speed up the Chinese job. We already have the past decade of experience. Thank you.

CYRUS NAMAZI: Thank you very much. Certainly I think ICANN will keep doing their best to facilitate the process of coordination between the three panels. I think if there's any clear need which you can identify in this context, I think it would be very good for each of the panels to pass that need to us and we will make sure that we do our best to address that.

Thank you very much, Wang Wei. This brings us to the end of the presentations. What you've seen is the work which is going on in the context of Generation Panels which are already formed. As you can see, even if a Generation Panel is not formally formed you can still get together, start the work, start the background work. We would still encourage you to keep in close touch with ICANN. We'll start with a proposal so that we can formally seat the panel. We are still looking for volunteers for many other scripts which are in the MSR. We're looking for volunteers to formulate Generation Panels and start the work for those scripts as well. If you are interested, please get in touch with us. I will share information about how to get in touch with us later as well.

Before we go there, let's see if you have any questions from the floor or any comments. Do we have any comments online? The Adobe Connect as well, if there are any queries there we can take those on as well. Any comments from the floor so far? Sure, Nicholas, please go ahead.

NICHOLAS OSTLER: I was fascinated watching the statistical comparisons. I'm talking to the Chinese in particular at the moment. I was very interested to see your presentation of statistical analysis of the incidents of characters in Chinese and Japanese newspapers, etc., but I was slightly worried about the relevance of that. I mean, what we're thinking about in setting up a

character base which will be adequate for the root is effectively giving the basis for putting in the names that you will find there.

The frequency of a character and whether it's all over the Internet or it's only occasionally seen in restricted journals is sort of irrelevant. You want a minimal adequate set and that's it. It seems a distraction to consider frequencies. I would appreciate some account of your thinking on why you've done these statistical analyses.

WANG WEI: Thank you. Basically, what we're doing, as I mentioned, is not the only one way and could not be the best way to doing that. Just suppose, to try to elaborate. If the mapping between Chinese community or Japanese community or Korean community, we need to figure out the most scientific way to solve the conflict.

My assumption is that if it might need to be agreed by all three Generation Panels. If the mapping doesn't need to be agreed by all three Generation Panels, it doesn't require any intensive work, like doing a statistical analysis.

Secondly, doing the statistical analysis is to just try to get a first impression how we're going to use this character, this word, in the real worlds and how often they happen to our daily life.

For example, if some characters could be quite intensively used in Japan and more closely but hardly used in China, probably we can both agree upon using a double rule that the Japanese propose. That's my intention. Again, I didn't say that's the only way and I didn't mention it would be the best way to solve the conflict. Thank you.

CYRUS NAMAZI: Do you want to come back to this point? Can we take somebody, any remote participation?

UNIDENTIFIED FEMALE: Hi, yes, I have a question and a comment from a remote participation. The first comment is from Will Tan. He says, "Continuing Andrew's earlier comment, I would like to note that the word "rules" in LGR encompasses (a) code points; (b) variants; (c) variant dispositions; and (d) WLE rules. While it's true that the Root needs a single, unified LDR, only (b) is effectively unified. (a), (c) and (d) are, as Andrew said, scoped to the script."

There was a follow-up question to that submitted by Yoshiro Yoneya. "(a), (c) and (d) will be different between languages which use the same script, so I think that it is not depend on script." Then he says, "Of course I understand in the Root Zone it doesn't have language context."

CYRUS NAMAZI: Thank you. Does anybody want to respond to that comment?

UNIDENTIFIED MALE: I'm a bit confused. It's not a unified LGR. This thing does not exist. There's multiple LGRs. We treat them, we process them as a word, but there's not a [singular] LGR. They are, in fact, separate. We don't really unify them on the [singular] XML file. That is not going to exist.

It's a [terminology] thing, but I'm a bit concerned where we use the concept of unified LGR because it does not really exist as a singular piece of data. It's basically a set of data that we do apply to various files

[inaudible] the prospective label can be using that. You have to process through all the different, various LGR that are created. You cannot really – in fact, it is not even possible to create a single XML file that will represent the intake rate or the various LGR coming from the different Generation Panels.

UNIDENTIFIED FEMALE: I think there's a response here from Asmuth. He says, "The LGRs are different by ISO 15249 script identifier. That is slightly different from Unicode script ID." Sorry, was that wrong? I don't know. I think he meant it in response to – okay.

CYRUS NAMAZI: Okay, could we take Michelle first? He was in the queue.

UNIDENTIFIED MALE: It was really about the concern by China. In fact, [inaudible] from China – is about the more restrictive rules going up the DNS tree, the fact that the connection between basically the second level Chinese character set and the one that you use in the root, there is not really a prescriptive limitations there. It's just that we're following SC 6912 which says when it's close to [point one] that basically the more you go up the tree, the more you tend to be restrictive.

That's kind of the guideline we'll be using if some Generation Panels think that they have to revise – the new character set or the new repertoire between the Root Zone and the current delegated second-

level zones, we don't really have a clear – it is not a requirement to be the same or a subset of the other one.

It's just that if you make a Root content that is bigger than the one that you already used on you CN TLD we're going to be surprised. We're going to be probably a bit more looking at why so that will create more questions obviously than if you make a subset. Basically you're going to say does that mean in the future your .CN is going to expand as well to be that way?

We would expect the .CN, for example, to be bigger. If it is smaller we're going to ask you, "Okay, that doesn't mean that you may have something else in the pipeline." It's more a question of surprise or a list to be consistent between the two. There is not a requirement, but again, we tend to follow the RFC to guide us in expectations between the existing TLDs and what you put in the root.

That's true not just for China, but it is true for every different script. We typically expect the root definition for a given script to be smaller, once you get that out to the respective TLD. Again, it's not cast in stone. That's case-by-case, and that [inaudible] things.

CYRUS NAMAZI: Thank you. Siavash?

SIAVASH SHAHSHAHANI: I am Siavash Shahshahani from .bazaar, new gTLD. My question is prompted by the comment of Andrew Sullivan. He's back here and could answer me. Big question is this. It seems to me that ICANN is

somehow requiring IDN TLDs to specify a language or language tag and I was annoyed very much when we were applying for the new gTLD. We were asked to specify a language for our TLD even though it was very clear that this was a script-based TLD, not a language based TLD. Finally we had to put down something, otherwise our application wouldn't have gone through.

The trouble with this is the following. It's obvious that something like the average script is used by a number of languages, and the same Label sometimes means the same thing, or different things in different languages, so it can be used. Seems to me that treating IDNs like this, requiring them to specify a language, is like treating IDNs as step-children of this whole domain business. So, I want to get an answer to that. Thank you.

UNIDENTIFIED MALE: You don't have to, by the way. It is a strong requirement for CJK because the Han script is basically shared between [inaudible] that are very different needs. So when you say, for example, Han, you have to say is it in Japanese context or Chinese context or Korean context [inaudible] context?

There's a lot of script, for example Greek, or even in some cases for Latin, someone to say, "Prove to me that you're going to [inaudible] rules for Latin." So it's not really a requirement, it's that for all scripts. Some scripts we need it, some no longer need such specific rules. Most of them, in fact, are not going need language specific rules.

SIAVASH SHAHSHAHANI: I would hope that in the next round of gTLD applications, ICANN will not require IDN applications to specify a language.

UNIDENTIFIED MALE: [Inaudible] in the context if we see a word, not in general gTLD. Sorry.

MATT STAFFBURG [?]: This is Matt Staffburg at .SA.

UNIDENTIFIED MALE: Sorry, can we maintain a queue if that's okay. Andrew is already.

ANDREW SULLIVAN: It's okay. I think it's the same topic.

MATT STAFFBURG: Right. I just want to clarify we work as a PVT service provider, and clearly with the new gTLDs you can have a Latin script table. You don't need to specify a language. You can specify a script instead. That must be a misunderstanding for the gentleman for .bazaar.

CYRUS NAMAZI: Thank you for clarifying that. Andrew?

ANDREW SULLIVAN: My name is Andrew Sullivan. Some of the confusion that seems to be going on here suggests to me that maybe people haven't read very carefully the procedure, because in fact, it's pretty crystal clear about

what this is. The first thing is that, on the topic of tags, they're, to begin with, optional. And secondly, they're not language tags. We went out of our way to expunge the word language all through that document for that very reason – because they don't have to be languages.

The point is that the tag indicates that this is a subset of code points that are used in more than one context. That's the only point of it. It's just arbitrary stuff. We happen to use a language tag because they're there and they're predefined, but they're not really languages.

One of the tags that is available to you under those circumstances is some sort of generic, or random, or something like that, which is not a language in any sense of the word.

The second thing is that it is important, in fact, to understand that there is a sense in which there is one LGR. That is, there is a single unified repertoire for the Root Zone and I don't want people to forget that. I'm nervous when I hear them saying, "We've got more than one LGR," because it's one repertoire. It's exactly one repertoire because it's one zone and that's all you can have.

The only thing that is important here is that you can use characters in the repertoire in sort of more than one way. That's what these tags do. They act as a shift bet. That's the reason that there isn't a problem with getting one repertoire.

So if the CJK community wants to spend some additional time outside of this meeting sketching on a whiteboard or something, I'm happy to talk about this in more detail, because really did go over this and that procedure should not lead to a big, complicated unification problem at

the end. You should have no problem there. I'm concerned that that seems to keep coming up. It's really designed to avoid that problem, because believe me, I was reamed about it several times while we were writing the procedure, so it was top of mind.

The final thing that I want to emphasize is that the policies for this zone are not necessarily going to carry to other zones. It's not quite true that it's either larger or smaller because it goes both directions. For instance, the .CA people (CIRA) have a policy and they allow a small number of Latin character, essentially the ones you need to write French and nothing else. Presumably .DE would have different set.

I would anticipate that the Root Zone would have, assuming that it expanded to include Latin characters, it would allow pretty much any of them, but it might not do variants at all. The Latin study, for instance, said no variants, it's just too hard to do, don't do it [in that thing]. So you might have a rule in the Root that says we're not going to do variants at all in the Root and if you want more than one label, it's two labels; just go buy it.

But that's not the way it works. In .CA, for instance, if I own color, I can also get C with a [CD] even though it's stupid to get – it's not a word, but it's a possible combination of characters, and therefore, I can have that.

I think that this is a key part of the flexibility of this design. That you end up with a system that can be reused in the machinery at various levels, but the policy doesn't have to be equivalent down the tree. The big problem with that, of course, is that it's going to result in difficulties for users.

The final point that I want to make, this is the reason that when people say "Well, there's this other character and it's also part of the language," the answer should be, "I don't care. Keep the minimal set. You want the set to be as small as possible for the Root Zone because you don't want to have weird characters that are going to function in strange ways in all the 2000 TLDs that we've got. That's going to be really confusing to people. After all, the whole point of this was that this be usable for humans. Thanks.

CYRUS NAMAZI: Any response to that? No? Thank you. Cary?

CARY KARP: I'm Cary Karp. I've been involved at numerous points in this process and have the distinction, I suppose, of holding the pen on the development of the IDN guidelines. Ever since the first version that was staff-drafted, which knew that there was some possibility of anguish by confusing the concepts of language and script, where the decision was made that everything will just be language. The document lacked clarity.

When script was meant, the word language was used and when language was meant, the word language was used. The effort was applauded massively when the distinction was made. In fact, the IDN guidelines refer to the way scripts manifest themselves in the namespace and although the consideration of language is obviously of fundamental importance when determining which code point repertoire one might want to put forward, nonetheless, that final act is enshrining the script repertoire, not the language that it's intended to represent.

Now I know that the confusion has returned. I'm not sure if I'm talking out of school, but I also specified the IDN component of the pre-delegation testing and note that many of the applicants have simply assumed that they're obligated to tabulate language repertoires and will submit a dozen language tables all sharing the same script, not understanding the individual languages, other than the one or two that they actually use to conduct their own business, causing a lot of difficulty for themselves not least, and presumably for the registrants as well.

All they would need to do is present a script-based table, and indeed in the policy statement to reflect a language-to-language variation in its application.

So it would be really wonderful if somehow during the remaining action what we really mean by script and what we really mean by language could be highlighted. Of course you start by considering the languages of the communities that are your constituency. But ultimately, a code point repertoire is just an ordered list of "This is what we want" and all of the operations that are performed on this are on the basis if Unicode properties, which script is the one. There is no language property there. You can tag and you can use taggings for policy, but again, we're talking about script here. We have to.

CYRUS NAMAZI: Thank you. Any further comments or questions? Is there anything on Adobe connect?

UNIDENTIFIED MALE: I hesitate to jump in where such authorities have just spoken, but perhaps some people are still slightly confused. The fact is that we are talking about scripts. The only case where there is a propensity to get confused I think is in the case of East Asian languages where, effectively, Chinese and Japanese and potentially Korean are traditionally written with a mixture of different scripts. In some sense, Hiragana and Katakana are different scripts, and they are different again from Chinese characters.

Nevertheless, what the decision has been made is that the set of characters which is used by Chinese out of all this lot is called the Chinese script. The different set which is used by Japanese is called the Japanese script; and similarly there would be a Korean script.

UNIDENTIFIED MALE: [inaudible] writing systems.

UNIDENTIFIED MALE: Yes, but I'm using script to [be] the same as a writing system.

That seems to me that that is where most of the confusion comes from. And if I could just get that particular innovative use by ICANN I suppose of the word script to mean the writing system as used by those particular languages. But otherwise a script is a script is a script. I think we're relatively clear. Cary will tell me I'm not.

CARY KARP: I long resisted the notion, but have subsequently come to realize that it has significant merit that the conditions that pertain in the ideographic

realm are different than the conditions contained in the alphabetic realm. But I can assure you that pervasive confusion is well-entrenched in the alphabetic realm.

UNIDENTIFIED FEMALE: We have another comment from a remote participant. It's from Asmuth, but he's still typing. He says, "Again, script is the ISO 5924 definition. It's not an ICANN innovation. The ISO 15924 definition does define Chinese and Japanese."

CYRUS NAMAZI: Any further comments on this? Any comments generally? Okay. So I'm going to skip the slides again and just jump towards the end here.

If you want to get involved, we have multiple e-mailing lists for you to join different Generation Panels, to apply for new Generation Panels, to get involved or to just sign up and see what's going on and provide your feedback less formally. Please come and visit the IDN TLD program web pages at the ICANN website. There are also links to documents which define the procedure, elaborate on the status of the project and also, of course, include outcomes and outputs of the project so far, including the MSR.

If there are no more questions, I'd like to thank all the panelists and the integration panel members, and of course all the audience for this interactive session. Thank you very much and let's close the session. Thank you.

[END OF TRANSCRIPTION]