SAN JUAN – IDN RZ-LGR Workshop
Wednesday, March 14, 2018 – 15:15 to 16:45 AST
ICANN61 | San Juan, Puerto Rico

UNIDENTIFIED MALE: Good afternoon. Session IDN RZ-LGR Workshop Open Session, Room 209-A from 3:15 p.m. to 4:45 p.m., March 14th, 2018.

PITINAN KOOARMOMPATANA: Hello, Sarmad. Can you hear me?

SARMAD HUSSAIN: Hello. Is this Room 209?

PITINAN KOOARMOMPATANA: Yes, Sarmad, this is Room 209-A in the Root Zone LGR Workshop Sessions. So we will start at 15. Thank you.

SARMAD HUSSAIN: Okay.

PITINAN KOOARMOMPATANA: He's on hold. Hi, Audric, could you please try speaking? We want to check the audio.

AUDRIC SCHILTKNECHT: Yes, this is Audric.

PITINAN KOOARMOMPATANA: Okay, perfect. Thank you.

AUDRIC SCHILTKNECHT: Thanks.

PITINAN KOOARMOMPATANA: Alright, so let's start at 3:15. So we are waiting for some more members. Thank you.

Okay, all. It's 3:15. I guess let's start. So everybody is welcome. Thank you for joining our session, IDN Root Zone LGR Workshop on 14th of March, 2018 – not 2017 as in the slide.

For these sessions, we don't have Adobe Connect, so for the ones who are listening online, if you have any comments, you can send an e-mail to IDN61-209A@icann.org. Amanda will be helping you reading out your comments. Okay. Next one, please.

So for the agenda today, this is the first session of the two sessions today for IDN. This one, we will go through the update from Integration Panels, then we will have some update on the toolset, and then we go to the community update which is the Chinese, Japanese, and Korean. We also have similar community updates for the next session as well, so if anybody who interested to see some samples, please stay tuned. Let's go to the first topic. Over to you, Marc.

MARC BLANCHET: Good afternoon. This is an update from the last ICANN meeting. The Integration Panel have been quite busy, actually, since the last meeting, and you will see what we've been doing. Just a reminder that the Integration Panel is a panel of independent experts. We have five people on that panel, myself, Michel here on my right, and three other people – Asmus Freytag, Nicholas Ostler, and Wil Tan.

We are tasked with reviewing the proposals presented by Generation Panels, and if accepted, integrating them into a consistent set of Label Generation Rules for the root zone. The decisions by the Integration Panel are required to be anonymous. One note is the Integration Panel obviously must take into account – that's actually a code from the document, the procedure – any public comments submitted in response to the posting of the Generation Panel output. Next slide, please.

We have reviewed a proposal for GP formations, the Sinhala. We have reviewed draft LGRs, some of them multiple times, multiple iterations, so we've been working closely with any Generation Panels: Cyrillic, Korean, Chinese, Japanese, Neo-Brahmi such as

Devanagari, Kannada, Gurmukhi, Gujarati and Telegu. By the way, those slides were done three weeks ago.

UNIDENTIFIED MALE: [inaudible]

MARC BLANCHET: Yes, and we have more – almost every slide here as additional stuff given it's been busy. So next slide.

We had a discussion regarding homoglyph with the Latin GP and also other GPs recently. Reviewed LGRs after their comment period, the Cyrillic. We have done some Han variants analysis, that is in fact a significant amount of work for the IP, especially my colleague on the right, because it's just many code points and many variants. Obviously, the other key factor in this is that there are various sources of data and usage, therefore sometimes the sets, especially for the variants, are not necessarily similar or identical. So it needs a lot of careful work with us and with the GP about this. But we're converging very well.

You may remember that up to now, we've released two LGRs for the root zone. The first one was also essentially a single script, and the others were multiple scripts, but they were very independent scripts. What we are having now coming is large

LGRs, script LGRs that are coming, and also the fact they have multiple cross-script variants, which makes yet another level of integration in terms of work and also make sure that the integration works fine, bug-free. So instead of waiting at the last minute, we started working on prototype integration of the big LGR's coming with the current draft version so that we could find issues and events. And when the actual integration will be done, then we will have cleared all the issues and events. Next slide.

We have issued a call for proposals for additional changes to the MSR-3. It was issued in November. We put it online on January 15. The idea here is that some GP needed some additional code points that were not in MSR-2 to be used in their own LGR, and therefore we wanted to make sure that the MSR on which they are based will include those characters.

The goal is obviously to do a minor update of this. We didn't handle any additional scripts, and essentially just addition of code points needed for some scripts. We received the following requests. I'm not going through the code points, but from Japanese, Latin and Chinese GP. Next slide.

The result is that we added three code points in Hani, three code points in Latin, and we didn't agree on adding the click in Latin, essentially because of security concerns, they're either non-PVALID or look at punctuation.

The public comments started on 15th of January, and deadline is 26th of February. Again, when we wrote the slides, the deadline was not yet done, so we're currently finishing the document with the staff for the summary of the public comments.

MSR-4 is currently not scheduled. It might include additional scripts when needed, for example, if there are Generation Panels coming with new scripts that are not in MSR currently or something, but currently not scheduled. Next slide.

Future work, I guess we may have more than what is on the slide, but on a high-level basis, we would like to produce a new Root Zone LGR, LGR-3. Obviously, it will depend on the delivery of the script LGRs after public comments and including whatever public comments happened. Target – which is just a target –Q3 calendar year this year. But we'll see how it goes.

Three groups of LGRs are being considered kind of coming: Han – so CJK – all the Neo-Brahmi and Latin-Cyrillic-Greek-Armenian, and again, these are kind of groups in a sense that we will handle them within the group, the group by itself, not separate scripts within the group to manage all the cross-group variants and the relations between those different related scripts.

However, there is a high probability that we will not create a single LGR with all those three groups, but instead, bundling a group per new version of the LGR. I think that's it for me.

PITINAN KOOARMOMPATANA: Okay. Thank you, Marc. Let's pause for a minute or two and take if there are any questions online or from the floor. Okay, so I guess we can move on to the next agenda, LGR toolset update by Audric. Audric, please.

AUDRIC SCHILTKNECHT: Hello. I will be presenting the update on the LGR toolset on the Viagénie team's behalf. Next slide, please. So the contents of this presentation will be quite short. We'll just sum up the LGR toolset in one slide and then move on to the new features that are planned to be released later this year. Next slide, please.

So those of you who don't know the toolset, the toolset is tool to abstract the editing process, the XML editing process of LGR documents. That means that people can focus on the creation of the updates, and we use LGR without having to edit an XML document. But also, the tool allows users to check things related to labels, like validating labels, generating variants, and also find out or check if there are collisions between labors in different LGRs or stuff like that.

The toolset is available both as open source – it's hosted currently on GitHub – and also as an online service. It has been developed as a set of command line and libraries in Python, and there is also the web interface which is kind of the main interface for the [inaudible] users. Next slide, please.

So I will move on to the updates to come this year. First, the new features. We will have an update to Python 3 because the tool is open source, it is available for the community, and we think that we should support the current version of the language it is written in and not being stuck to the old version. We will also add MSR-3 once it is finally completed. I think one of the biggest features which is kind of a visual improvement is to be able to handle very large LGRs, for example CJK. And so we are working on improving the overall performance of the tool to be able to deal with the largest LGRs.

We will also add a new function to check if multiple LGRs are harmonized. That means that we will check that the variant code points are transitive and symmetric in each of the LGRs selected and that the variant sequences in one LGR cannot be nonvariants in another LGR. Then we have some kind of less important features which are more improvement to the interface. For example, when we validate labor, we will add a specific rule which fails if the label is not valid. Also, we'll try to improve the error message when the syntax of the XML document is not correct. And one of the new features is also to add the variant code points when you add a new code point from another script. Next slide, please.

Continuing on the UI and UX improvement, we'll have a button that will allow to populate every symmetric and transitive variant missing from the LGR. So you'll just click on the button and it will add all the missing variants so your LGR is well formed. We will add a new tag page which will allow people to manage code points – I mean the tag assigned to each code point, but from the tag, instead of going through each code point and adding a tag. And we will do kind of the same with WLE rules.

There will be a new function to populate the LGR. You can select the scripts and then it will automatically add the range of code points from the script into your LGR document. We are also trying to improve the summary. Currently, the summary is kind of text output, so we will make it look a bit nicer. Also, we will rename it as validate because it's actually more than just a summary. It also ensures some of the properties of the LGR, like for example, that all your code point variants are symmetric and transitive.

And then we have a bunch of updates to the HTML output. For example, we will in the HTML output display both the number of members and the number of mappings for variant table. Currently, we only display the number of mappings. We will add a function to make links contained in the references. We'll try to reduce the number of lines from the variant tables because it's actually a bit doubles. And yes, we will also have a new function that allows you to input a label and will display the free form of labels. So the A-label, the U-label and the Unicode sequence. Next slide, please.

There are of course some bugs to be fixed. For example, currently we cannot change the variant types, so it's kind of a broken [block]. Also, the default type has been updated in the [inaudible] for the LGR. It is [validated] from block to block. And this has not been reflected correctly on the interface, so that will be fixed. We will fix from the statistics, for example, the tag counts on the HTML output will be fixed. We'll also add to the HTML output the number of code points contained in the largest set.

And finally, we'll fix some issue in the tools. For example, the annotation tool does not display the first label. So we will investigate that. Yes, so this was a comment from the last ICANN session that variants were not validated using the LGR, so this will be fixed. So that means that when you generate a variant, it also should be a valid label in the LGR. Next slide, please.

So as I was saying, the release will be later this year, so we are planning a June release. And some links where you can find more information about the toolset or how you can use it with the user manual. I think that's all from me. Yes. Thank you.

PITINAN KOOARMOMPATANA: Thank you, Audric. Any questions from the floor? Okay. Please, Dennis.

DENNIS CHANG: So this LGR tool, I believe – it doesn't say explicitly, but I believe it's an implementation of RFC 7940.

PITINAN KOOARMOMPATANA: Yes, it is the requirement.

DENNIS CHANG: Thank you. So the follow-up comment and question is – I'm also a sitting member of the ICANN IDN Guidelines Working Group, and in there, the current working draft will require registry operators to implement RFC 7940 to publish IDN tables. So I would assume that at that point in time, some registry operators will use these LGR tools in order to produce those tables. So it will be useful to make the experience – or give registry operators assurance that this tool is actually compliant with RFC 7940 so they don't have to go back and do that analysis themselves. Thank you.

MARC BLANCHET: You're looking for who to say that? because it's open source, right? It claims conformance to that RFC. But you need more?

DENNIS CHANG: No, I haven't seen the actual source, so I don't know. Right, maybe in the release notes. If it says so, I think that would be sufficient, right? It's just here in the explanation overview, I didn't see explicitly – yes.

MARC BLANCHET: Well, I think it's like the fourth presentation of it, so we try to remove the basic stuff.

DENNIS CHANG: Understood.

MARC BLANCHET: One thing that may be appropriate for the operators of current IDN tables is the fact that the tool also enables you to import the previous RFC 37-something into the tool right away. So you could essentially convert. And that's the web interface, but all the functions of the web interface are also available as libraries or command line, so if people prefer typing, then...

DENNIS CHANG: Yes. Thank you.

PITINAN KOOARMOMPATANA: Mats, please.

MATS DUFBERG: Is the goal that the tool should test everything in the RFC to make sure that it's fully compliant with RFC? So like dates are checked, etc. Today, the current version does not, as far as I know.

MARC BLANCHET: We have implemented multiple test cases, but obviously, the spec is pretty large. Having full test coverage will be a project by itself.

UNIDENTIFIED MALE: But I think that is really important, that ICANN really provides something that fully tests the tables, as long as it's possible to test, of course. But things like dates and such fields are possible to test, and I think that should be included in the tool.

PITINAN KOOARMOMPATANA: Okay. Thank you. Very good suggestion, so we'll take that back. It's actually developed based on the [inaudible] is the following RFC, but explicitly says it complies with completely [inaudible] of the test. We'll take that back and we'll get back to you. Okay. Alright, any more questions? Anything online? Okay. So I guess we can… [inaudible] please.

MARC BLANCHET: I think we had that discussion about dates, and if I remember well – I may be mistaken, or it's maybe the version – I think I remember looking at the RFC and it was like defined as a string, therefore you cannot verify. Maybe the version. The version was a string. People were expecting 1, 2, 3, the number, and it's actually a string. So therefore, there's nothing you can test. I don't know about the date, but –

UNIDENTIFIED MALE: [inaudible]

MARC BLANCHET: Oh, okay. Well, whatever.

PITINAN KOOARMOMPATANA: Okay. Alright, so I just want to repeat, anyone who wants to submit a comment, e-mail address is icann61-209A@icann.org. Alright, in the meantime, let's move on to the community updates. So for the first one, may I invite Chinese GP Chair to share the status?

WEI WANG: Thank you, Pitinan. Thank you, everyone. My name is Wei Wang, Co-Chair of CGP. Another Co-Chair is Kenny Huang from Taiwan. He couldn't attend this meeting in person, so I will help to introduce the updates of CGP. Next, please.

The CGP has 23 expert members from ten countries origins, including China, Taiwan, Hong Kong, and Macau, where the Chinese script was set as the official script, Singapore, Malaysia, as well as members from Europe and also America.

Luckily, we have Edmon Chung as the advisor to help us to coordinate as consultant and help to coordinate [various] IP. And during the process, we work together closely with J and K to define the repertoire, the variant set and the final XML LGR. Also, we have the IP consultation to help us. Next, please.

Since 2014, we have generated over 10 versions of the proposal. The repertoire changes a lot, and the [inaudible] changes a lot. And luckily, as the IP introduced, we expect to close the panel in the third quarter this year. We are all working hard to achieve this goal. Next, please.

The issue about the CJK is because the Chinese characters were broadcasted to Korea and Japan, so we have a lot of overlap of characters in our repertoires. Next, please.

As you might see that within the CGP, we defined a repertoire based on the CDNC table and the DotAsia table, and also, we take some official document which defined and normalized the Chinese character into account. Next, please.

And also, we have many overlapped characters with J and K, and we need to coordinate with them on these characters and their variant mappings. Like two years ago, the Korean community proposed that there are about 258 unacceptable variant groups, so we spent about one year talking about this issue, and finally, 146 variant groups were affected and changed. Totally 445 variant mapping entries were affected. Next, please.

The variant definition of Chinese domain labels is characters with different visual form but with the same pronunciations, and it was the same meanings as corresponding official forms in the given language context. As you might see, there is an example about the Chinese characters and their variants. Every code point in the CGP repertoire has its own preferred simplified variant and the preferred traditional variants. These are supposed to be allocatable, and the others are supposed to be reserved or blocked.

A code point might have a reflexive, preferred and simplified and traditional variant, and a code point might have no reserved variant. So accordingly, the CGP defined a subtype of allocatable type and block type based on the definition of Chinese simplified and traditional variant. Next, please.

And also, we have some out of repertoire variants. That's because it seems we have many overlapped characters with J and K, and there might be some characters which were used in Japanese environment as independent – or not independent, but as Japanese Han characters. Might need to be revealed by the CGP expert to see if the variant mappings should be reset considering that the Japanese character was imported. So 42 characters would be considered as out of repertoire of variant characters.

And there are another two characters in the CGP repertoire which were not included in the MSR-2 but which have been added into the MSR-3 for which I appreciate that IP accepted the request to add them. Next, please.

And there's another big issue in the CGP proposal, is that there are 136 characters with multiple allocatable variant mappings, so which will lead to multiple allocatable labels. And the number is not fixed number. Theoretically, if the lens of the label is not enough, we might have numerous variant allocatable labels. To address the issue, we analyzed the 136 variant sets one by one and trying to eliminate the multiple allocatable variants. Next, please.

So we defined some new subtypes to reduce number of allocatable variants, and it is proven that this way could be reduced number of allocatable labels. Finally, the number of allocatable labels for the CGP will be under four, which means at most, we will have four allocatable labels. Next, please.

And also, we have the latest feedback from IP about to clarify the need to include the TGSCC which means normalized Chinese characters published by the government in 2015, and to show if it is really needed for the domain name scenario and to provide more reference information to the characters and to the variant mappings using available sources.

[Our] review the variant sets which differ from the second level practice between CDNS and DotAsia. So there might be about 50 variant groups might need further investigation and review. Next, please.

So for the next steps, we will keep communicating with IP on the remaining issues, and we will update the proposal to include Unihan as reference source for code points and variant mappings, and we need to reevaluate whether the 18 TGSCC characters and the 14 imported Japanese character are necessary. We might remove them from the current repertoire, which will make the CGP repertoire exactly the same as the second-level practice of DotAsia.

After that, we will reduce the repertoire and variant mapping tables. And KGP LGR proposal might be affected because we remove them and we will redefine the variant mappings related to these 60 characters. And in this meeting, we've run into another issue that is about the [various] similarity of variants which means in Japan community or the Korean community, the kana, kanji, hanja and hangul characters might be treated as visually similar variants to each other, which will make the [inaudible] a little bit more complicated. So we are trying to address this issue with K and J together.

Anyway, we will start our review process on the variant groups which have different variant mapping solutions in CDNC and DotAsia. The review work might be done by the end of next month, but in a joint meeting of CNDC and CGP. Thank you. That's all.

PITINAN KOOARMOMPATANA: Thank you. Are there any questions from the floor? I don't see anything online. Okay, so let's move on to the next item on the agenda, Japanese GP updates. Hiro Hotta-san, please.

HIRO HOTTA: Thank you, Pitinan. The update from JGP, Japanese GP is very brief, I believe. So as we're still old GP so we have four steps. Populate JGP with diverse experts, define the requirements and basic framework for Japanese LGR based on the expertise and experience of Japanese IDNs for 20 years. And coordinate with C and K GP.

Step three has been almost done, as Wang Wei said, but maybe something remains regarding the variant definition between hanja and hangul in K LGR, and maybe in Japanese script between kanji and hiragana, katakana there may be a variant definition. If there is, we have to go back to step three again. But it's under investigation. And step four, finalize LGR following necessary consultation with IP and Japanese community. I'll talk about this step four today. Next, please. Yes. Next, please.

The first version of Japanese LGR. First version means that only consider the Japanese situation and not the Chinese or Korean LGR. And as you see in the variants section, for kanji, Japanese LGR defined no variants for ourselves. But final Japanese LGR would import variants of Chinese LGR and Korean LGR. So we are very passive in this sense. Okay, next, please. Next, please.

Reduction of the number of allocatable labels. The big issues in front of us are two of them. One is the reduction of the number of allocatable labels, and second one is the variant definition between kanji, hiragana and katakana cross-script variant. The first one is reduction of number of allocatable labels.

So as consulted with the IP, JGP is trying to solve by limiting allowed strings by employing the notion that allocatable labels basically consist of daily use kanji. There are around 2000 kanji characters among 6000 repertoire characters. And case two is a daily use kanji plus kanji that's [intended for] personal names.

Case two was proposed by some of our Japanese community, but after I wrote this slide, the case two may be diminished. So I believe we can choose case one, only that they use kanji. This means that the variant labels which consist of only daily use kanji can be allocatable, and others are locked. That's what this says. And it significantly reduces the number of allocatable labels among the possible variant labels.

And as I said, JGP is considering case one works fine in reducing the number of allocatable labels. XML of case one has been developed for inspection by IP, and we submitted such XML just one week ago. So I think IP is inspecting that, and maybe IP needs more information why this XML is written like that. So it's being inspected by IP, I believe. Next, please.

And the second big issue is handling similar-looking characters, especially cross-script ones. In Japanese writing systems, any combinations of Japanese characters are used to express Japanese words. For example, the trade names or trademarks can be a string in which hiragana, katakana or kanji characters are anywhere in the string. Even the ASCII character can be anywhere in the string. So very flexible string can be a Japanese word.

So the initial intention of the definition of Japanese labels was there are no variants in Japanese LGR except those imported via variant definitions of C and K. However, IP proposed us that some of them looking similar, they should be handled as variants. So we are inspecting that, and so far, the kind A punctuation characters 30FC and 4E00 and 30FD and 4E36, it may be better to define being variants, because they are a single stroke and punctuation.

And kind B, this is a harder proposal. Mutually resembling kanji characters and kana characters such as katakana and kanji, maybe those other than Japanese people looks – they are similar, or even identical. But from Japanese eyesight, they are different. And other katakana and hiragana, they are somewhat different. You can identify it.

They're proposed to be variants, and the kind A is considered acceptable, but kind B is not acceptable. That's our first response. But we are still under investigation whether they can be or they should be defined as variants. Okay, I think this is the last slide. Okay, thank you.

PITINAN KOOARMOMPATANA: Okay. Thank you, Hotta-san.

UNIDENTIFIED MALE: Can you please go back to the slide? Back. Yes, this one. So I don't understand the characters of Japanese characters, but when you have given these three examples of three variant strings, these variants are because of similar looking, or sounding similar?

HIRO HOTTA: Sounding similar and meaning similar.

UNIDENTIFIED MALE: [inaudible]

UNIDENTIFIED MALE: Okay. And second question. Next slide. Are these punctuations allowed in TLDs to be registered?

HIRO HOTTA: I'm sorry.

UNIDENTIFIED MALE: Punctuation marks in TLD.

MARC BLANCHET: Yes, this is a PVALID so they're perfectly fine. They're not really punctuation, per se. It's a prolongation sound so it's not a [inaudible] categories of [inaudible] character.

HIRO HOTTA: They can be part of a word.

UNIDENTIFIED MALE: Yes, that makes sense now. Okay.

PITINAN KOOARMOMPATANA: Okay. We have Mats. Please.

MATS DUFBERG: A question to IP. Does IP think that the kind of B, that they should be treated as variants in Japanese?

MARC BLANCHET: That's a loaded question. Obviously, when they look exactly the same, saying that they're unique doesn't really get you anywhere, because if the font uses the same glyph to display both characters, saying that they're unique doesn't really save you from safety issues or security issues, because you have basically two strings that look the same encoded differently.

Obviously, there's this homoglyph, as we know is not an exact science sometimes because sometimes true for every font, sometimes it's not true. Like obviously, you will see more differences in serif fonts than you will see in sans serif. Like in the sans serif – this is using sans serif here on the screen – [does] anyone see a difference between the katakana and the kanji? They look exactly the same. It's not true in every font. So there is an open debate.

We're not drawing any conclusion here, but the debate and the discussion needs to take place for the extent of those – and I would hate to use the term visual similarity, because that's really a term that is dangerous, because obviously by the Root LGR project, the procedure is crude visual similarity from the Root Zone LGR. We are not supposed to – because that's supposed to be done by a higher process. But those are not, in our opinion, a visual similarity. We are talking about homoglyph here, basically things that look the same for most people, even including in Japan.

UNIDENTIFIED MALE: The first pair in B, for me, I don't see the difference here. But the second pair –

MARC BLANCHET: Yes, that's why there is judgment to be made here. And the list is not finished, the list is not defined at this point. It's basically being studied. I think there was a debate obviously on what is the list of those homoglyph if you want within the Japanese writing system. Yes, these are just examples. I agree with you that the first pair looks much more confusing than the second one. But there is way more than that. You could create at least 15 of them that look the same.

UNIDENTIFIED MALE: I'm trying to apply this discussion on Latin generation too, and I find that the discussion is different time, so I get very confused, I have to admit.

MARC BLANCHET: To some degree, I think the C, J and K panels are now discovering that beyond semantic variance which [inaudible] focus for C, J, K panels they were really more concerned about, especially in the Chinese context between traditional and simplified where the characters look completely different but they mean the same thing. So it's a semantic variant. In fact, also in Ethiopic, there was kind of a similar thing where characters look different but they mean the same thing or sound the same. So each LGR is own set of to some degree context on this, so the answer is different in some cases.

But what is really not different, if you have homoglyph, these are to be treated very seriously in the LGR. There's no way that we can accept to have identical code points that don't have a variant relationship. I don't think that's acceptable no matter what. Then you can develop beyond that for semantic or even phonetic variants. That's possible. Each LGR [inaudible] GPs is free to develop variants that go beyond homoglyph, but then they have to make a case for it.

The case has been done pretty clearly for C, J and K for semantic variants. Again, we've got the same issue with Ethiopic at some point, we had to [inaudible] variants that were going beyond homoglyph. But in the minimum, you have to take care of homoglyph. That's always on. To some degree, maybe it was a surprise for the C, J, K panel that they also had to deal with homoglyph, but it's not a new thing. In fact, this issue had existed for a long time. For example, in the second-level LGR for Japanese, in fact some of us did work on that. There are in fact some visual variants in the second level for Japan, so it's nothing new.

Obviously, there you have to draw a line. For example in the second level, people working on it did limit themselves to simple characters, so we kind of look at simple strokes, characters, but to some degree, that's a judgment call. At some point, you have to decide of how far you go on that slope, because it's a slippery slope. It's not exactly – and as we have seen now with Latin, we know that there are some homoglyphs that are – it's not a complete black and white case.

UNIDENTIFIED MALE: And the kind A, do you consider that to be two homoglyph pairs? 30FC and 4E00 and 30FD and 4E36 respectively.

MARC BLANCHET: I think that's a case on the second level for Japanese at this point. I have to look at the definition, but I think that's the case. Obviously, there is some other case that sometimes vertical positioning is slightly different or the length, or sometimes some subtle differences depending on the font, so you have to kind of look at them – it's not like the case where in Latin, Greek or Cyrillic where you have absolutely perfect homoglyph. In this case, it's never that perfect.

UNIDENTIFIED MALE: What do you mean by second level? Second level domain?

MARC BLANCHET: I'm referring to work that was done for the second level LGRs by ICANN. There was work done for Japanese, Chinese, Korean, on some European languages. That was done, completed last year, I think. And some of us worked on it, so we have some expertise on that too.

UNIDENTIFIED MALE: [Thank you] for the information. The kind A, it's in the proposed second level LGR reference, in the current version. So they are defined as variants.

MARC BLANCHET: Just proposed, by the way. It's out there. It's just a reference so you don't have to use it, but it's a reference.

PITINAN KOOARMOMPATANA: Next we have Bill. Please.

BILL JOURIS: For kind B, perhaps it would be useful to say all of these are considered unique by users of Japanese. But would they be considered unique by someone who only speaks Chinese? Because that's the sort of confusion that could be a problem. Someone who only speaks English isn't going to assume he has any idea what's going on, but someone who speaks Chinese and goes, "Those look like they're the same to me" even though someone who speaks Japanese would go, "Oh, of course they're different." Just a thought.

PITINAN KOOARMOMPATANA: Okay. Professor [Kim], please.

UNIDENTIFIED MALE: Thank you. [inaudible] In the middle, there are case one and case two. I want to check if I understood the situation correctly. When you say by limiting allowed strings, are you talking about the variant labels, not the repertoire itself?

UNIDENTIFIED MALE: [inaudible]

UNIDENTIFIED MALE: Right. So [inaudible] to change the repertoire but –

UNIDENTIFIED MALE: [inaudible]

UNIDENTIFIED MALE: I see. Okay. Thank you.

UNIDENTIFIED MALE: Well, I think – go to the next slide. If I heard you correctly, that you are saying these kind A are the characters which become part of a specific label. They are not usually used as a punctuation itself. Is that [variant] understood correctly? Is that correct?

HIRO HOTTA: Yes, they are used within a word.

UNIDENTIFIED MALE: Within a word. So it may be a good idea to remove the word "punctuation" from here, because it is making to me that punctuations are used to break the sentences, and there are no sentences in a label.

HIRO HOTTA: Okay.

MARC BLANCHET: Just sound marks and not punctuation.

UNIDENTIFIED MALE: Yes. So just the characters. That's good enough, I think.

HIRO HOTTA: Yes, people may misunderstand it because of the word "punctuation" [in a sense].

PITINAN KOOARMOMPATANA: Okay. Any other questions, comments? Alright, so let's move on .We still have time, so we'll come back to some other questions at the back. Let's go to Korean Generation Panel updates. Dr. Kim, please.

KIM KYONGSOK. Thank you. I'm Kyongsok Kim and I make presentation of K LGR. Next, please.

In K LGR, you have two scripts: hangul and hanja, and the Korean script usually means hangul. However, in the context of K LGR, Korean script refers to a union of hangul and hanja. Okay, next slide. Next, please.

In K LGR, for the 1.0 that was published in December last year, and it has 11k Korean hangul syllables and there are no variant groups. And there are 4758 hanja characters, and there are 152 variant groups. In addition, there are five variant groups composed of hanja syllables and hanja characters. And out of those five, three hanja characters are out-of-repertoire variants.

The hanja character set is composed as follows. It is the union of two sets. One is KS X 1001. It has 4620 characters. The second set is IICORE K column. It has 4743 characters. And when you make union of those two, it becomes 4758. Next, please.

Unification of variant groups for hanja between KGP and CGP proceeded, and both GPs reviewed three of four Chinese variant groups. Those contain two or three K hanja characters. K hanja character means hanja character included in K LGR hanja repertoire. The result of unification of variant groups between KGP and CGP are shown below.

K LGR version 1.0 has 152 variant groups, and each of those variant groups contained two or three hanja characters. The other Chinese variant groups were split so that no more than one Korean character in Chinese variant groups. So currently, there's no conflict in variant groups between K LGR and C LGR when you [concede] variant groups composed of hanja characters only. Next, please.

Here is brief history of KGP activities. Next.

In January this year, KGP sent K LGR for public comment, and the public comment will close soon. It is March 17th. And then a summary report will be given to KGP March 24th, and KGP will probably modify K LGR based on public comments, and we'll send it to IP and then IP will evaluate the proposal. And if everything goes fine, then it'll be integrated into subsequent version of our Root Zone LGR, hopefully. Thank you.

PITINAN KOOARMOMPATANA: Thank you, Professor Kim. Any questions or comments? We don't have anything online as well. Okay. So I guess we can open the floor for other comments as well. Can you go to the next slide? Another one. Okay. So that's the wrap-up of the sessions.

UNIDENTIFIED MALE: Sorry.

PITINAN KOOARMOMPATANA: No, we're not going to close, I'm just going to find out how to get more connected to the program. So ICANN.org/IDN, and then also the e-mails. So if you have any other question after this, you can follow up as well. So now, please follow the questions.

MARC BLANCHET: A follow-up for Mats. I think the way the toolset was done was that we wanted people to use it as a loose editor of the LGR. Therefore, at the time of entering the data, you could enter a lot of different things, and we validate the data. You could enter a lot of different things, and we validate after. And in the validation, the actual dates are validated.

The idea here was like source code, right? With an editor, with a source code, you can do wrong stuff, but then you compile when you're done so the actual – that was the philosophy inside the tool, is to accept more, accept very liberally what people will be entering, but then validate afterwards so they can work as they want. It's actually dated, our dates are validated at the time of validation.

PITINAN KOOARMOMPATANA: Okay. So any other questions, comments? We actually have half an hour for discussion as well. Please feel free to use the time if any questions from other IP. I think that will clarify us a lot of things as well.

WEI WANG: Actually, I'm preparing some slides for tomorrow's meeting with IP especially on the visual similarity issue. We had a discussion this morning within CGK, and after the meeting, I have a brief introduction for the discussion with the [inaudible]. I was wondering if… I can understand the rationale of why we raised that issue of similarity variants. I was just wondering if we need to – I was wondering if the disposition on variants and the similarity variant should be a little bit different.

I give Pitinan a scenario that when someone, the first applicant applies for a Latin label, and at the same time applicant B applies for a Cyrillic label, which the two labels are visually similar labels. When we apply at the same time, I bet [they] must go to some [disposition] period to address the conflict.

But similarly, what if someone applies a kana label with Japanese script but for the Chinese community, the Chinese community users have no idea someone in Japan is applying a label which will affect their future application. If it happens to be some Chinese guy applies for some Chinese label with visual similarity of the kana at the same time, they must go to some disposition process. But if they don't apply at the same time, so they lost the chances to avoid the risk.

So I was wondering this morning when I discussed with Pitinan if we need a little bit different extra process to handle these visual similarity labels, which might be some – I call it warning period or some – set it not as allocatable or blocked but alerting to let the community get involved in this, more community members get involved in the label process. That's a rough idea. I haven't thought this through, but I think I need to think it over tonight and discuss with IP tomorrow morning.

PITINAN KOOARMOMPATANA: Okay. [inaudible] please come in. Then next, Dennis.

MARC BLANCHET: If I take the case of Cyrillic on Latin, for example, it's really first come first serve because each of them are blocking the other one. So if someone applied to, let's say, a label that looks exactly the same as the Cyrillic one, it will block the other one. So it's basically the first one, either the Cyrillic one or the Latin one, the first one that was applied for will win. It's no different from any other situation, except obviously the scope is a bit wider because you have to pay attention to the integrity of the LGR.

And integrity of LGR will give you that answer right away. When you apply to a new label, if you use the RZ LGR, you will see right away that you have a conflict, and then you'll be blocked. So you have to basically pay attention to the Root Zone LGR that is published, because you will show that in the content of the data that this is going to be blocked.

I understand it may be a surprise that you could see, for example, a situation where hangul sequence who could in fact block a Chinese label completely, but that will be obviously very visible in the Root Zone LGR dataset itself. So this is nothing new.

To decrease a bit the list for the kana, for example, on the CJK characters, it's not that common – the way I know – I know a bit of Japanese – you don't really put random Kana in the middle of kanji. There's kind of an order of things you do in the label even in Japan. Typically, you don't put kana in the middle of kanji. So it's not completely random, it's not a mix of random things. Typically, you would see either a bunch of katakana together followed by some kanji, or at least not totally arbitrary. Don't mix and match like that.

But yes, there is a risk. Obviously, if you add – I don't want to use again "similar" because we're not talking similar. Similar is outside the scope for Root Zone LGR. We're talking identical, things that would be – I mean you could look at the definition of variants in the procedure. I can link to you, but it's not defined as visually similar. It's not. I want to be clear on that. It goes beyond that, because if you're just talking visual similarity, that's done by another panel that decides, "Oh, these are really too close, we shouldn't allow this one to be delegated" or whatever. So it's beyond us. We're not dealing with visually similar here, we're dealing with homoglyph, things that look the same for most users of this, including native from the country. So it's not that, "Oh, they just look kind of the same." No, they have to be the same for most people.

PITINAN KOOARMOMPATANA: Okay. Dennis, please.

DENNIS CHANG: Thank you. I just want to respond or make a comment to your second case where an applicant applies for a top-level domain name and that might block future needs of a different organization, and how the application process works. So I'm not an expert, I'm not speaking for ICANN, I'm just speaking from my experience, for my company's experience going through the application process in 2012.

So you apply for a TLD and there's a process whereby ICANN do as much publicity as possible through all of us, the community, in order to let everybody know all of us, the community, in order to let everybody know – or as much as possible, again – that there is a process, an open window to apply for TLDs. And these

TLDs' applications are posted for public comment. So there is a chance for other individuals or organizations to either support or object to those applications.

And that window, it was a long window, and so somebody apply for one and he thinks or the organization thinks that it's not suitable for a TLD, they can follow the process, object or try to block the application for any arguments. Either is trademark or visual similarity or whatever. And [Michel] is right, for visual similarity items, there is a review process for similarity review, and they will provide the steps for these applications to move forward, either through – one has to withdraw from the process or they have to go to auction or some sort of relief.

PITINAN KOOARMOMPATANA: Any other comments, questions?

DENNIS CHANG: So I just want to take the opportunity that we have the IP and fellow Latin GP folks here. Bill, if you want to come here perhaps. So in the Latin GP, in the repertoire, we finalized our code point list, right? And there are two code points, the schwa (Ə) and the turned e. And I have done a little bit of research on those characters, and by the strict definition of homoglyph, they're not homoglyphs because the upper cases are different. Is that correct or not? We need to do more? Because if that's the case, then our variant analysis would result that these are not variants because they're not homoglyph because of this unification by case property.

MARC BLANCHET: Yes, if you go with uppercase for variants between Greek, Latin and Cyrillic, you're opening a giant can of worms of things that will make so many characters [inaudible] each other that you couldn't even create – that will even block ASCII labels. So that's kind of probably why people aren't even considering it, because it would in fact block – you would have a situation where now you will have pure ASCII labels blocking each other. I don't think that's acceptable by anyone that you could have suddenly TLDs in ASCII blocking each other.

In fact, you may even have a situation like currently allocated TLDs could in fact be invalid per those rules. So I don't think we want to go there. I understand there is limitation there. You could argue that that's not good enough to – but we look into it. In fact, we did consider some case [inaudible] but it's not [workable]. I don't see how that could even go further. And I think one of the main reasons is because you would be blocking ASCII TLDs. So you can't even think of it.

DENNIS CHANG: But that was not the issue here. The issue here is that we have two code points that are homoglyphs as lowercase, the turned e and the schwa. So there is no difference on them in lowercase. But they have different uppercase appearance. They look completely different uppercase.

WEI WANG: Okay. Well, I did not understand that way. So essentially, we ignore uppercase because that's out of scope of IDNA.

MARC BLANCHET: It's not [inaudible] the front. That's the same issue.

DENNIS CHANG: So turned e and schwa are identical in lowercase, and reasonable conclusion since we ignore uppercase is that those two are homoglyphs and should be variants. Is that...?

MARC BLANCHET: Yes, that's correct.

UNIDENTIFIED MALE: [inaudible]

DENNIS CHANG: Thank you.

WEI WANG: That's correct, yes. Sorry.

DENNIS CHANG: I mean there are other cases where the uppercase are identical in –

UNIDENTIFIED MALE: Yes, for most letters.

DENNIS CHANG: In Latin. But here, we're talking about lowercase identical.

UNIDENTIFIED MALE: Lowercase behavior, yes.

PITINAN KOOARMOMPATANA: Okay. So let's see, nothing from the online. Okay, so maybe I'll put myself in the queue and others do have time to do so. So just to summarize what I've had a chance to discuss with Wang Wei, Chinese GP just now is that – so right now, let's say somebody apply for katana two characters which have also the same in kanji. And the thing is the final objective that he wants is also two visually identical labels shouldn't go to the different owners. So this is I think the common ground here that that will create confusion for the users.

Then how to make all of us aware that if somebody applies for those katana, then the one who use kanji should be alert and come and take a look. That's where we have to define the variants. Otherwise, we wouldn't have known that these labels regenerate another set of possible looking similar that will be blocked. So I guess maybe we gradually have more common understanding after more studies.

MARC BLANCHET: One thing I would like to add too is that for example, pretty much every script – not CJK, but all the other ones – use some form of the letter O, a circle. If you go to the absurd, you could create obviously a label constitute of multiple Os and it'll be confused everywhere. There is a limit on what you can do. You can obviously create a katakana level. It'll look kind of stupid and very long, but it doesn't make sense in either Japanese or even in the translated version of Chinese, because it doesn't make sense.

So obviously at some point when – this is a mechanical process. Root Zone LGR is a mechanical process, so you will get some result out of it. But there will always be some eye on top of it that will say, "No, this is looking stupid and shouldn't be delegated." The mechanical system can only allow so much because there's some sort of AI if you want on top of it to make sense of it.

Otherwise, like I said, the multiple O example is a good one. We did not for example try to – I know at some point people were trying to propose, "Oh, we should mix some Hindi variant system" because that's something that look like O, so suddenly you'll make an O from the Hindi and South Asian languages that's confusable with the Latin O. You could go pretty far on that slope on being basically lost.

BILL JOURIS: Plus you can get an O and a crescent or a C and a straight line or an L. You get those in lots of scripts.

PITINAN KOOARMOMPATANA: Okay. Hotta-san, please.

HIRO HOTTA: Thank you. So I'd like to make a question or request to IP or ICANN. Is there some [stance] that identically stroked or identical characters must be variants? If there's such requirement, it's easier for us.

MARC BLANCHET: Yes. Frankly, I wish there were, but there is not, because first of all, it's a variable target. We found in fact as time goes with modern evolution of France, we're finding in fact the phenomena that France, that we present different culture are getting what they call harmonized, and suddenly you're getting more variants than you used to have. I've seen that in Armenian.

Armenian was a recent example, so a newer France that covers, let's say, Latin and Armenian have suddenly created way more variants than we used to have because people want to create an Armenian document, they want to look like – they want to be able to mix, if you want, in the same text Latin – doesn't mean on the same line or the same word, but in the same text. You know, both languages. And they want to look – both of them look good together. Doesn't mean you're mixing inside a word both languages, but you can put them on the same – let's say you're doing a translation. We've seen that happening quite a bit, and it's disturbing because obviously it's good for a layout or typography design, but it's deadly from a confusability point of view.

We have seen that even in Japanese. We've seen that in some newer font or UI fonts where they're making really romaji and Japanese characters looking nice together, and so you have this kind of harmonization. So we can't really define the table. People [inaudible] Unicode as a table of confusable, but to some degree, it's so loose that it's rather useless because everything is similar to each other. In the end, you get so many cases of false positive and confusability that it's not useful. So there are multiple people, multiple sources for that, but there's not a single source.

Frankly, sometimes the best source is usual fonts you find for OS or platforms. Then you know your best font, [bigger] font, or whatever, you get the thing inside and you see what's the reference and use the same thing.

So yes, I'm sorry, there is no bible for that or reference that you can use. You have to kind of do judgment on – that also makes the IP work kind of difficult, because even for us it's difficult sometimes to make a judgment.

HIRO HOTTA: Maybe the definition of identical or not is not easy to do that. I understand that, but for example, identical characters must be variants. Is there any statement like that?

MARC BLANCHET: Where we define them as homoglyph. Homoglyph is, again, let me take the example of Greek, Cyrillic and Latin because it's an easy one. In those, in fact typically the same font in every platform use the same glyph. It’s not even copied, it’s the same glyph that’s referred in the cmap table which is inside the font system. It’s mapping between the code point and the glyph, and in fact you will have the same code points from each of the script will map to the same glyph in the font. It’s very common, it’s very easy to see.

First you see that in the text, but then you can even check that on the font itself, that multiple code points are in fact using the same glyph. That’s a clear case of total confusability. In fact, nobody can make any difference between the same string written in Cyrillic and Latin because there is none. There are absolutely no differences for those characters, they’re exactly identical.

HIRO HOTTA: So homoglyphs must be defined as variants.

MARC BLANCHET: Absolutely.

HIRO HOTTA: Okay. Where is it written?

MARC BLANCHET: I think the best thing you can find – let me go – in fact I was looking at my text. Okay. I'm just going by the procedure. So an IDN variant as understood here is an alternate code point or sequence of code points that could be substituted for a code point or sequence of code points in a candidate label to create a variant label that is considered the same in some measure by a given community of Internet users. That is the definition we're working with. That's [currently] the procedure for the Root Zone LGRs. That's the only thing that is really totally defined. All we have [to live] is that. That's because we don't have better than that. On [inaudible] we use that definition in the procedures to basically say that homoglyphs are clearly encompassed by that definition.

PITINAN KOOARMOMPATANA: Please.

HIRO HOTTA: I didn't hear in that definition that it's a requirement to be homoglyphs. It could be some differences.

MARC BLANCHET: Absolutely. That way semantics are covered. Semantic traditional simplified are covered by that.

HIRO HOTTA: But also visual differences.

MARC BLANCHET: Well, then let me read another piece of the procedure. That's in section B 3.4.2. Finally, investigating your possible variant [inaudible] Generation Panel should ignore cases where the [inaudible] is based exclusively on aspect of visual similarity. That's [inaudible] I'm living by the procedure, so every time I have a doubt – that's why I don't like the term "visual similarity" because we're not supposed to deal with that.

UNIDENTIFIED MALE: Can you please read that again?

MARC BLANCHET: Yes, if you look at the procedure, that's a document that rules of IP, basically. It's in section B – like boy – 3.4.2. There are three references – I look at it visual similarity in the whole document. They're mentioned in fact three times. The third one is the more important one, I think, the one that I just said. I'll say it again. "Finally, investigating the possible variant relations, Generation Panels should ignore cases where the relation is based exclusively on aspect of visual similarity." But visual similarity, it just looks the same, not identical. So obviously, we're playing a bit United Nations here, we're playing with words.

HIRO HOTTA: But isn't homoglyph an extreme case of visual similarity?

MARC BLANCHET: You could say that, yes.

HIRO HOTTA: So we should ignore homoglyphs?

MARC BLANCHET: Well, if you do that, then obviously, we'll have opinion to the contrary. So obviously, we are interpreting a document here. You could interpret different ways, I guess, but that's why there is an IP, I guess.

PITINAN KOOARMOMPATANA: Okay. Wang Wei, please.

WEI WANG: Just follow-up Hotta-san's question. Actually, when I got to know that there's kind of these identical homoglyph issues, I was a little bit confused because in Chinese repertoire, there are more Chinese characters look so similar to each other than kana and kanji. So why we didn't get this kind of feedback from IP before? If we have a definition for identical homoglyph, there might be hundreds and thousands of cases in Chinese repertoire. But why we just give the examples between kana and kanji or hanja and hangul?

MARC BLANCHET: Come on, Wei, you're pushing it here. If the character – we did the same, they would have been unified. There's a unification process for [10606 Unicode] to make sure that characters that are unifiable have been unified. So I understand sometimes the differences are pretty subtle. Sometimes even there is depending on the sources between Chinese – you could have a complicated situation where a character looks the same as another code point because in a difference source, the unification works a different way.

So there are some places, some specific cases you could argue that. I agree with that. Maybe you have to explore them. So that's also why having a bigger repertoire makes things a bit more complicated, because suddenly, you may have – because difference with CJK that the characters look different between the different sources. You may have Japanese characters and Chinese characters that have the same Unicode code point, but in fact they look kind of different for the some code point. There is in fact some situation where you have two different code points with two different sources and the visual in fact kind of cross between the two code points.

So you could argue if you're using different fonts, let's say you're using a Japanese or Chines font, you could have in fact similarities that go across code points because the sources use different reference glyph for both code points. You have quite a few of those. I don't think there are thousands of them, there are a few of them. I know from memory of working on those things that there's a few of them.

The good new is that you only have 20,000. You don't have like 80-90,000, that kind of number we have in CJK now, but we have [enough.] But at least it's really big. We have 19,000 and some change, so you would probably find maybe a few dozen. I don't think it's more than that. You're probably talking about a few dozens of those cases where you have unification going across code points with that situation. I could find probably some [inaudible] quickly some of those. But for that, yes, you have to look at the multicolumn display and see what's going on. They tend to be pretty close to each other so you can see what those are. So it's not a simple thing. I'm not saying it's simple.

PITINAN KOOARMOMPATANA: Okay. We probably have two minutes left for the last comment or last question. Okay. Then we'll have conclude [inaudible].

UNIDENTIFIED MALE: When the discussion is over on LGR, I have a question a little beyond LGR with these three people.

PITINAN KOOARMOMPATANA: Okay.

UNIDENTIFIED MALE: We haven't used that [inaudible] 15 minutes.

UNIDENTIFIED MALE: Yes.

UNIDENTIFIED MALE: Can we have that?

PITINAN KOOARMOMPATANA: So let me just conclude this.

UNIDENTIFIED MALE: Yes, we can actually have that. Yes, we can have that next session.

PITINAN KOOARMOMPATANA: Okay. Thank you.

UNIDENTIFIED MALE: Thank you.

PITINAN KOOARMOMPATANA: Alright, so thank you everybody who joined this session, Root Zone LGR Workshop. Now the session is closed and we'll be back at 5:00 p.m. Thank you.

UNIDENTIFIED MALE: Good job.

[END OF TRANSCRIPTION]