ICANN IDN Variant TLD Program Meeting Minutes

Second Face to Face Session for Project 2.1:
A Way to Develop and Maintain the Label Generation Rules (LGR) for the Root Zone in Respect of IDNA Labels

DAY 2: Saturday, 13 October 2012

Meeting Chair: Dennis Jennings
Discussion Leader: Andrew Sullivan
Meeting Minutes: Dennis Chang
Meeting Location: Toronto

Participants:

In Toronto: Alexei Sozonov, Joseph Yee, Chris Dillon, Andrew Sullivan, Dennis Jennings, Asmus Freytag, Michael Everson, Neha Gupta, Akshat Joshi, Syed Iftikhar Shah, Zhang Zhoucai (Joe), Alireza Saleh, Daniel Kalchev, Dmitry Kohmanyuk, Rinalia Abdul Rahim, Nadya Morozova, Edmon Chung, Panagiotis Papaspiliopoulos, Yoshiro Yoneya, Mirjana Tasic, Vladimir Shadrunov, Francisco Arias, Nicoleta Munteanu.

Remote participants: Dennis Chang, Sarmad Hussain, Raymond Doctor, LinLin Zhou, Naela Sarras

1. IDN table development process steps discussion: an example diagram presentation to be incorporated into the document as appendix.
    a. Identify variants
    b. Identify set of characters
    c. Complete IDN table
    d. Apply IDN Table (i.e. Label generation)
    e. Applied for string: calculate IDN Variants (permutation/substitution)
    f. Disposition
2. Discussion of the same using real example in Chinese

Primary Codepoint		Preferred SC	Preferred TC	Other Variant Characters	
U+53D1 发	U+53D1 发	U+767C;U+9AEE	發;髮	U+5F42;U+9AEA	彡;髮
U+5F42 彡	U+53D1 发	U+767C	發	U+9AEA;U+9AEE;U+767A	髮;髮
U+767C 發	U+53D1 发	U+767C	發	U+5F42;U+9AEA;U+9AEE	彡;髮;髮
U+9AEA 髮	U+53D1 发	U+9AEE	髮	U+5F42;U+767C	彡;發
U+9AEE髮	U+53D1发	U+9AEE	髮	U+9AEA;U+5F42;U+767C	髮;彡;發

3. Example of different rules under Chinese/Japanese tags

Tag: Chinese

Input	Allocate	Withhold	Note
戀 U+6200	U+6200, U+604B		Traditional Chinese character of U+604B
恋 U+604B	U+604B, U+6200		Traditional Chinese character of U+6200
爱 U+7231	U+611B, U+7231		Traditional Chinese character of U+611B
愛 U+611B	U+7231, U+611B		Simplified Chinese character of U+7231

Tag: Japanese

Input	Allocate	Withhold	Note
戀 U+6200	U+604B	U+6200	Considered Old Form of U+604B
恋 U+604B	U+604B	U+6200	Current form of U+604B and U+6200
爱 U+7231			THIS CHARACTER DOES NOT EXIST IN JAPANESE. This is the simplified Chinese character of U+611B
愛 U+611B	U+611B		There is no rule defined for relationship between U+7231 and U+611B

OUTPUT:

Base	Allocate V	Withheld V	Blocked V	Tag	(source)
戀 U+6200	U+604B			“Chinese”	C p1
恋 U+604B	U+6200			“Chinese”	C p1
恋 U+604B		戀 U+6200		“Japanese”	J p1
爱 U+7231	U+611B			“Chinese”	C p1
愛 U+611B	U+7231			“Chinese”	C p1
愛 U+611B			U+7231	“Japanese”	p2

Application for: U+604B, 611B

Variant strings generated:	Chinese case:	Japanese case:
a) 604B, 7231	a) Allocatable	a) Blocked
b) 6200, 611B	b) Allocatable	b) Withheld
c) 6200, 7231	c) Allocatable	c) Blocked

4. ISO 15924. For Japanese we can assign script tag for Hiragana and Katakana or language tag.
5. By having tag repertoire, we are getting logical tables or code points but ultimately they are all part of the overall repertoire. Withheld code point may not be in repertoire.
6. Reference: Unicode Standard Annex #29: Unicode Text Segmentation www.unicode.org/reports/tr29/
7. Simpler solution could be one Han table combining Japanese and Chinese but the issue could be that Chinese will have to deal with much of characters that they will never use. Using language as divider could be as simple. Similar cases exist with Arabic as well.
8. Two views: Keep one table but allow the concept that some characters are only used in Chinese or Japanese. Or two lists for each language.
9. When you register, you will need to select a tag.
10. At the end, these are code points, irrespective of script because it is one root. Secondary panels will be responsible for being aware of other script work.
11. The primary panel should be aware of the secondary . Primary panel should be doing the most generic work it can.
12. We do not want to build in rules that are dependent on linguistic rules. No spell checker like rules. Need to keep it simple and generic.
13. We will have to rely on the fact that no one will pay quarter of million dollars to register nonsensical string. However, there may be cases that someone will register a very popular string look-alike. - Visually similar string.
14. The process to develop LGR can use language specific data. The disposition of resulting labels needs that information too.
15. Secondary panel can generate a block variant in Japanese based on variant relations in Chinese. Further discussion of item 3 above.
16. There should be a tool that helps the secondary panel with this so that they don't have to work by visually comparing the characters.
17. Further discussion using variant relationship using Chinese and Japanese: Application for: U604B, 611B (see item 3 above). Walking through the example.
18. Language table or script table? Does it matter? Script table could be simpler.
19. We could let applicant decide which variants they want. Choice by applicant would make it simpler.
20. One table for Chinese. One table for Japanese. Each tag would have its own table.
21. If someone else comes along and ask for withheld, the answer is no. The original applicant could ask for withheld and be granted.
22. Blocked case would be useful for visual similarity case.
23. User expectation is the fundamental reason for the variants.
24. DNS security would be another reason for variants.
25. .asia example: practical case: Chinese Han set was superset of Japanese Kanji. Almost 20% going to variant today.

26. However, we should note that there are Japanese Kanji characters that were never in Chinese.
27. We must get it in the document if this is important that a certain disposition case is a decision such as “Must be allocated.” However, there is no way to enforce the “Must be allocated.” In the same sense, we may not need “Withheld” either. Only blocked or not. Not feasible to enforce allocation. ICANN is not in position to enforce. Possibly by contract to a certain degree?
28. Primary panel treatment is invisible in secondary panel output.
29. We need to be able to separate policy issue from technical issue.
30. User Experience project (P6) will be producing guideline or suggestions for gTLD registry contract. The enforcement by contract is out of scope of this project (P2.1).
31. What if applicant wants variants in multiple languages? Use script tag rather than language tag? In another word, script tag only. Language tag is a specialization of script tag. It’s hierarchical.
32. Would it be permitted to apply using only Arabic tag without specifying language tag in current scheme? Yes.
33. Advantage of having a language tag could be for blocking.
34. Tags we are using are already defined in RFC 5646. UND=Undefined is also used.
35. For the purpose of developing this language tag could be useful.
36. Unicode script property, we’ve abandoned as one of our criteria. ISO 15924 script subtag is used.
37. If we do not find reason for language tag, we should delete it from requirement. If we receive input with public comment, then we could restore it.
38. Most radical simplification could be that we do not use blocked variant either. If an applicant makes a mistake, they could come back and ask for the correct one later. If no harm could be identified, we should not have additional rules and keep it simple as possible.
39. There might be a need in Arabic for language tag. But might be isn’t a strong enough reason for inclusion.
40. For scripts that are not easily grasped on the whole, it is actually in violation of our principle if applicant asks for everything in the script.
41. Rather than excluding from whole, build from subsets upward. This technique could fit our conservatism principle better. Sub-repertoire.
42. Two cases of confusion:
    a. Running text domain name that depends on application’s interpretation.
    b. Two labels that confident users can confuse the two labels.
43. Recommendation: Draft should be modified to ask for UND-xxx. Not be language specific. Not include in this process for now but make provision for future addition.
44. Recommendation: Block only based on comments received to date.
45. Language tag is already subset of script. Also, it’s designed to be more restrictive. Purpose for this is that no one gets around the rule by claims of language specific. That would undermine the general case.

46. Primary panel provides sub-repertoire and secondary panel either accepts or rejects.
47. Primary is script based. Primary Panel is responsible for one or more script or each panel only one. These are two options.
48. Block should be default unless there is strong reason to allocate. Such as the simplified and traditional Chinese. Primary panel would need to provide justification if not using default.
49. What happens if primary panel first submit based on sub-repertoire and then later find out an issue with a larger repertoire?
50. Latin panel originally recommended no variants unlike CJK where variants are well understood.
51. The request for more details on how we know when primary panel is ready is valid. However, we have to leave it to the panel to use their judgement. Also, there is secondary panel who will ask questions. The risk is there that the secondary panel will miss important question. It's doesn't seem possible to come up with a full formula to know that the panel is doing their job well.
52. This model is the best we have so far. Bottoms up with expertise.
53. Is the group satisfied with this model? That we depend on secondary panel to determine if the primary panel has done a good enough job.
54. If more text is desired, Andrew would appreciate written text provided him. B.2.1.3 section is an attempt to address this.
55. Keep in mind that the Goal of secondary panel is to turn it down if they have doubt. Conservatism principle here. We need to ensure we write this process so that they are disinterested party. They also need to be unanimous.
56. There are also expert advisors to keep the primary panel in line.
57. It would be much more useful to provide sample cases. Rather than trying to come up with algorithm.
58. Criteria may be useful to improve judgement without trying to come up with hard and fast rules. Principles we have here does that somewhat already.
59. Recommendation: to add more in terms of criteria rather than going down algorithm.
60. Pakistan example: more than 50 languages. Started in 2008 in multi stakeholder committee. Three technical committees: Language, Policy, and Keyboard. The language table formulation process. Decided and agreed to go with single language table first. Compromised to create single language table. Many of characters had similarity and confusability: direction of dots for example. This process took 3 years. There are more emotional and aggressive but we should focus on the main purpose. Because we are in initial stage, there are many things we don't see. Initial hypothesis are proven false later.
61. Maintaining the LGR. What could be issues 3 years from now? There will be new code points. Unicode isn't done. Little concern that if we have to repopulate after a long period of time, we don't have the primary original core panel so that they will have to reinvent and redo the same work already done. Less concern with secondary panel. No suggestions but a concern.

62. Big cases with large populations will be finished. But for instance for Latin, new communities which are currently under represented on the internet will come to the floor. We will need to integrate new body of people and political pressure to accommodate new request. There will be pressure to integrate Indigenousness languages in the internet to promote use.
63. Another view is that 3 to 5 years, we'll still be in the thick of things. Unicode is an example. The panel would likely find other missions to continue.
64. If Applicant Support Program does what it is supposed to do then we may find that we have smaller community participate. ICANN may need to do better job in outreach.
65. We should be looking to see if there is living language community. For practical reasons. Perhaps "Used in the DNS" should be added to the requirement. This may be more of a policy requirement. So we should stay away from this.
66. Protocol, not policy should be place to make limitations.
67. How do we prioritize secondary panel work? Probably not a big issue.
68. Issues List from yesterday:
    a. Timeframe: continuity of experts for secondary panel concern. The secondary panel could be essentially full time job initially. Pent up demand would be there. Section on risk in the report could be useful.
    b. Cost Issue: how many panels? Supporting them will be costly also. Advisors will cost. How will this fit into ICANN organization? There will be a first discussion tomorrow. No answer at this point but there are future projects that are being planned.
    c. 12 scripts applied in current gTLD so we'll need panel to support these.
    d. Secondary panel could be 3 to 5 members only.
    e. We'll need to fall back on public comment period.
69. Panel expertise gap could be addressed in many ways: advisors, current practitioners, etc.
70. Criteria could be useful such as regional representation.
71. ICANN should be make DNS advisor available.
72. Explicit inclusion principle. There is risk that subset will have conflict with later larger set. This is where advisor will be helpful.
73. Based on goodwill could be risk. Should be self-correcting because the default will be no.
74. Weighing factors will work better than checklist because they rely on judgement. This will guide the output without boxing those who will need to make decision.
75. Primary panel is not a representative party. Instead it's a work group. The point of Public Comment process is exactly so that despite the fact that the panel may not representation is full. The secondary panel is expected to tell the primary panel to fix it if community comment asks for such.
76. Involving GAC and ccTLD could be an option.
77. Ultimately, judgement and public oversight is what we are talking about.
78. Communications between primary and secondary panels – what is the concern? The current document is informal. Request for more text in what needs to be

provided to secondary panel from primary panel. Formal proposal and reply format would be helpful and will be written in.

79. Reply time deadline by secondary panel would be necessary. Secondary panel could advise the primary pane how long after reviewing the submission.
80. ICANN staff is limited and advisors will also be limited. So primary panel start needs to prioritize somehow. First come first serve could be an option such as DNS.
81. Overall program management status reporting needs to inform what is going on to all interested parties.
82. First mover advantage issue: may or may not be a big issue. Not sure how to address this in the document. Be factual but not hysterical?
83. Interaction between results of this process and new gTLD program: Current round is being evaluated without this being done. Again, not sure what to say about this in the document. What Applicant Guide Book says is that the variant will be blocked. IDN gTLD would be likely to get into the root before we are done with our work here. But don't know what will happen. There is mechanism such as objection.
84. There is a chance that we will discover a variant issue we have not yet considered. Probably a slim chance but not impossible.
85. No Greek gTLD application in this round. But it's possible that some application could cause conflict for future Greek.
86. Next version of document expected to be out for comment? Depends on how it goes with comments and Andrew's time. Will advise.