ICANN79 | Forum de la communauté – Prochaine série du programme des nouveaux gTLD : séance de travail pour planifier une revue de similarité de chaînes
Mardi 5 mars 2024 – 3h00 à 4h00 SJU

PITINAN KOOARMORNPATANA : Cette séance va commencer, veuillez lancer l'enregistrement.

Bonjour, bienvenue à cette séance de révision sur la similarité de chaînes. Je suis Pitinan, je serai responsable de cette séance qui va être enregistrée et qui va être gouvernée par les normes de comportement de l'ICANN.

Pendant cette séance, les questions et les commentaires seront lus s'ils sont envoyés dans la partie des questions et des réponses de Zoom. Si vous souhaitez poser une question ou faire un commentaire dans la salle, levez la main sur Zoom. Nous vous donnerons la parole et vous pourrez allumer votre micro.

Nous avons un service d'interprétation pour cette séance qui va inclure l'anglais, le français, l'espagnol. Donnez votre nom et la langue dans laquelle vous allez parler si vous ne parlez pas l'anglais. Lorsque vous prenez la parole, parlez à une vitesse raisonnable.

Je donnerai maintenant la parole à Sarmad.

SARMAD HUSSAIN : Merci beaucoup.

Bonjour, bonsoir à tous ici et en ligne. Cette séance va présenter les travaux que nous avons faits concernant les orientations liées à la révision des similarités de chaînes. Voici d'abord le contenu de notre présentation. Il s'agit d'un travail qui fait partie de ce qui va concerner la nouvelle série de nouveaux gTLD. Nous allons parler de chaînes TLD, de la similarité des chaînes et de la façon dont on peut faire la différence entre les différents types de chaînes.

Il y a un PDP accéléré dont la phase 1 a demandé ce que l'on élabore de nouvelles orientations concernant la similarité de chaînes. Ce processus est en cours de réalisation. Il est réalisé de manière transparente pour les candidats et pour la communauté. Et ce que nous faisons dans le cadre des recommandations de ces lignes directrices au niveau des SubPro et au niveau du PDP de sa phase 1, nous avons commencé à élaborer ces lignes directrices.

Il y a deux problèmes que nous sommes en train de résoudre. Le premier est le fait que nous devons déterminer de manière claire les mécanismes permettant de déterminer ce qu'est une similarité de chaînes pour que tout le monde comprenne. Le deuxième défi que nous essayons de résoudre actuellement, la deuxième difficulté, concerne les TLD variantes et l'espace dans lequel cette similarité de chaînes a lieu et augmente considérablement. À cause de cela, il peut être très difficile de réaliser une révision complète de ce problème.

Nous proposons un processus prévérification mécanique, quelque chose qui serait basé sur des données que nous allons collecter à travers la communauté. En fonction de ces données que la

communauté nous fournira, nous allons réaliser un processus de révision de ces chaînes et de leur similarité. Nous avons une série d'étiquettes sur lesquelles nous pouvons travailler déjà.

Ce que nous avons fait jusqu'à maintenant représente la première phase de ces lignes directrices qui sont actuellement présentées aux commentaires publics. Vous avez le lien ici de ces lignes directrices. Nous avons aussi une série de points qui peuvent être utilisés pour la partie de la transparence et pour les processus dont j'ai parlé pour aborder ce travail de manière plus efficace.

Nous avons travaillé avec une équipe qui travaille sur ces lignes directrices. Je vais maintenant donner la parole à Michel Suignard qui est un boursier d'Unicode qui a travaillé avec nous sur ces lignes directrices et qui va nous présenter ces lignes directrices. Si vous avez des questions, vous pourrez les poser sans aucun problème.

MICHEL SUIGNARD : Merci. Je voulais aussi dire que j'ai participé au panel d'intervention du côté de l'ICANN. J'ai travaillé sur le LGR de la zone racine qui provenait du panel de génération qu'on devait intégrer. J'appartiens à ce petit groupe d'experts. J'ai une certaine expérience concernant les variantes, comment travailler dans un contexte multiscript dans la zone racine. Les variantes sont très importantes ici. Nous le verrons dans cette présentation.

La première chose, lorsque de nouvelles applications sont analysées, nous devons analyser les variantes pour qu'elles ne soient pas

éliminées, les variantes qui font partie d'un processus en deux étapes, une présélection et un processus d'examen.

Nous allons analyser quelques concepts pour une question de terminologie. Vous savez ce qu'est cette possibilité de confusion et de similarité des variantes. C'est quelque chose qui a l'air d'être la même chose. Ce peut être le « c », ce peut être le « ç », en arabe il y a aussi une lettre. Ce sont des formes qui sont semblables, qui se ressemblent, qui peuvent être éliminées dans la zone racine. Dans la deuxième ligne, vous voyez « ç », sigma en grec, vous voyez aussi des exemples en devanagari. Ce sont des choses qui sont similaires visuellement seulement, ce ne sont pas des variantes. Dans la dernière ligne, vous voyez des choses qui sont visuellement distinctes, ils ne sont pas vraiment similaires visuellement.

La première chose, cette capacité de confusion, c'est un degré auquel deux étiquettes peuvent être substituées l'une à l'autre sans que l'on s'en aperçoive. Les utilisateurs ne savent plus quelle étiquette choisir. Quand on utilise aussi la confusion, c'est-à-dire quelque chose qui peut être confondu, c'est lorsque deux étiquettes possèdent un degré de confusion qui dépasse un seuil défini. Ce seuil est une question de définition parce que ce n'est pas vraiment blanc et rouge. On va définir ce seuil.

Ici, nous allons parler de zones racine et il y a certaines contraintes. Ces chaînes peuvent être très courtes. La zone racine n'est pas encore prête à recevoir des chaînes à deux lettres. On peut en voir quelques-unes ; dans le cas du chinois, on peut avoir des situations dans lesquelles on a

un seul caractère et parfois, cela a l'air d'être une seule syllabe, c'est une variable. Cela inclut les domaines ASCII. C'est quelque chose de tout à fait différent que nous faisons pour la zone racine. Nous ne considérons pas les domaines ASCII, nous ne considérons pas les variantes au sein d'ASCII.

Cette similarité de chaîne est une apparence de choses que l'on ne voit pas. Il y a par exemple des lettres qui changent en fonction de la police. Un « i » cyrillique peut ressembler sous sa forme italique à un « u », le « f » de nouveau lorsque vous avez sa forme italique, on a une lettre qui ressemble à un « f ». Dans le LGR, on a éliminé ces deux possibilités. On voit parfois la frontière entre la similarité visuelle de leurs différences, et le panel de générations pour la zone racine a décidé de les notifier comme étant similaires.

Les variantes sont différentes dans la mesure où on essaie d'établir ce qu'est une variante comme identique ou similaire. Un niveau de variantes va être le même mais ne se limite pas à la partie de similarités visuelles. Ce peut être la même chose au niveau visuel. On peut avoir un seul niveau ou plusieurs niveaux. On peut aussi avoir une apparence visuelle qui se rapproche, mais aussi au niveau du sens et au niveau de la sémantique. Les cas les plus difficiles de sémantique sont les cas du chinois. On a des variantes qui passent du chinois simplifié au chinois traditionnel et à ce moment-là, on applique la phonétique et on ne fait pas cela dans la similarité de chaînes. La similarité est seulement visuelle, on ne parle pas de la sonorité.

La même chose pour les équivalences de variance. Par exemple, si on lettre trois lettres, A, B, C et si E est équivalent à B et que B est équivalent à A, à ce moment-là on a des termes inégaux entre A, B et C et la disposition des variantes est aussi en général un bloc, ce qui signifie que les niveaux de variantes ne sont pas autorisés et ils peuvent être alloués. Cela veut dire qu'il y a des délégations multiples qui vont utiliser le niveau de la variante, etc. On applique des règles très strictes pour limiter le nombre de situations. Cela s'applique au chinois, à l'arabe surtout ; ce sont les cas les plus typiques de ces situations où l'on va allouer une variante. C'est pour cela qu'on utilise le terme de niveau original, niveau primaire ; c'est ce que l'on considère comme l'étiquette de variantes.

Il y a beaucoup de différence entre le niveau original et le niveau de variante. Il faut examiner tout cela dans les études. C'est ce que l'on a fait, parce qu'il n'y a pas vraiment de différence claire entre les deux origines de ces variantes. Par exemple, vous voyez ici dans le cas du « F », c'est exactement le cas.

Lorsque nous faisons cette étude de la similarité, nous devons faire une tâche de reconnaissance et de dépendance parce qu'on ne peut pas évaluer la similarité. Il faut évaluer d'autres questions aussi. D'abord, il faut voir quel est l'utilisateur cible et le contexte. Il y a des réponses différentes ici dans ce domaine de confusion possible. Si vous regardez le script d'abord, il vous faut analyser le script et connaître la langue. Il faut avoir une certaine familiarité avec le sous-ensemble ASCII parce que la zone racine et ASCII dans sa majorité ou presque, 80 % ou 90 % de la zone racine se fait grâce à des chaînes en ASCII. Voilà le contexte

dans lequel on doit reconnaître ces deux niveaux. Les deux ne sont pas obligatoirement présents en même temps. Il va y avoir un niveau un et dans d'autres contextes, cela va ressembler à celui que vous connaissez.

On peut aussi avoir une approche entre langues ou entre scripts, c'est-à-dire différentes langues qui utilisent le même script. On peut ne pas reconnaître les similarités de scripts que l'on ne connaît pas. C'est différent, mais ce n'est pas une exigence. On peut vous demander de reconnaître des différences au niveau du chinois, par exemple des deux types de chinois, mais c'est quelque chose qui se fait au premier niveau, au niveau de la chaîne.

À nouveau, nous n'essayons pas de lire un texte de littérature. Nous essayons de reconnaître les identificateurs. Ce sont des chaînes de lettres ou des chaînes de scripts que vous voyez qui peuvent se ressembler, qui peuvent porter à confusion. Il y a un concept holistique qui vous permet de reconnaître tout cela. C'est bien et c'est moins bien parce que parfois, votre cerveau ignore les différences parce qu'il ne les attend pas. Nous verrons tout cela un peu plus en détail.

Évidemment, pour la révision, vous verrez ensuite les concepts. On a d'abord l'étiquette candidat, c'est celle que vous demandez dans la zone racine. Ensuite, on aura les étiquettes déjà déléguées et leurs variantes. C'est important parce que quand on compare, on ne compare pas qu'aux étiquettes déjà déléguées, mais il faut comparer aussi avec les autres variantes. C'est important parce que les variantes ne sont pas seulement visuelles. Si on prend le chinois par exemple, il

va falloir comparer celles qui se ressemblent évidemment, mais aussi celles qui ont la même sémantique, même si elles ont l'air différentes. C'est important que ces similarités soient prises en compte également.

Ensuite, on a aussi une série de mots réservés. C'est un mélange d'acronymes et de noms. En général, c'est pour les ONG ou d'autres organisations. Vous ne pouvez pas mettre de restrictions. Aussi, les étiquettes appliquées peuvent porter à confusion et être confondues avec d'autres noms. Quand on fait une demande pour une nouvelle étiquette, vous ferez partie d'un grand groupe d'étiquettesf appliquées et c'est là que vous faites partie de la même étude, parce qu'il faut comparer toutes les étiquettes appliquées ensemble. Cela fait partie de cette étude de confusion. Et la cible de révision, c'est en fait un petit ensemble d'étiquettes qui devront être comparées sur le plan visuel avec l'étiquette candidate et pour la cible d'examen. Voilà pourquoi c'est considéré individuellement ainsi qu'avec toutes les autres qui ont fait l'objet d'une demande. Prochaine page, s'il vous plaît.

Quel est notre objectif ici ? Il faut veiller à ce que toutes les étiquettes soient au moins considérées pour une délégation et qu'elles soient suffisamment différentes pour éviter les confusions. Bien entendu, on ne se fixe pas un but de 100 % ; il arrive toujours que certaines personnes sombrent dans la confusion et on va voir pourquoi dans les prochaines pages. Il faut vraiment se concentrer sur celles qui prêtent le plus à confusion. Ce n'est pas parfait, on vise à prendre les meilleures décisions étant donné les contraintes auxquelles nous nous heurtons. Prochaine page.

Maintenant, comparons la similarité aux variantes, parce qu'on utilise le système de variantes pour résoudre notre problème dans une certaine mesure. Je vais me répéter par moments, mais soyez indulgents, je veux vraiment que ce concept soit bien compris.

Encore une fois, les variantes sont les mêmes. Quand vous regardez l'espace, c'est censé être au même point alors que la similarité, il y a toujours ce qu'on appelle l'espace de perception. Ils ne sont pas vraiment au même endroit si vous regardez ce qui est similaire. Par exemple, un cas clair, c'est quand vous avez un triplé de caractères similaires. On en a deux qui peuvent ressembler au dernier, mais pas avec celui du milieu. Celui du milieu ressemble aux deux extrêmes, mais les deux extrêmes ne se ressemblent pas entre elles. C'est une situation qui arrive assez souvent, donc c'est ce qu'on appelle la perception.

Encore une fois, les variantes ne sont pas limitées à leur similarité visuelle, ce qui pose encore d'autres difficultés, comme je l'ai dit tout à l'heure. Et les variantes ne tiennent pas compte du tout des majuscules ; il faut en parler, alors que dans la similarité visuelle, il faut tenir compte des majuscules et des minuscules, et cela pose d'autres difficultés aussi. Bien entendu, les variantes ne suffisent pas à faire des évaluations de similarités visuelles. Ce n'était pas le but. Le but, c'était de trouver ce qui était identique et pas juste similaire. Les variables n'apportent pas de réponses à toutes les situations de similitudes.

C'est ici qu'on voit ce qui pourrait nous aider à créer ce qu'on appelle un ensemble de variantes. C'est ce qu'on utilise dans le système LGR pour créer ces règles sur les dispositions établies. C'est en fait un

mécanisme où vous avez un ensemble et aussi une cartographie entre chaque point ou chaque séquence pour déterminer ce qu'il faut faire avec. Et on utilise le système LGR. D'ailleurs, je ne veux pas non plus vous donner tous les détails, cela ferait trop.

Pour ce qui est de la grappe d'éléments qui peuvent être confondus, c'est un ensemble de tout ce qui peut être confondu. On les regroupe par variantes. Encore une fois, vous pouvez créer une situation transitive. C'est le problème, on pourrait créer trop de chaînes qui peuvent être confondues ; c'est pour cela qu'il faut faire attention. Il faut expérimenter un peu plus de ce côté. Encore une fois, cela pourrait être utilisé en prévérification et en postvérification L'idée ici, c'est d'utiliser le système des variantes pour nous aider. Prochaine page.

L'interaction avec la similarité de chaînes et les variantes. Quand on le compare à l'ensemble délégué, il faut tenir compte de cet ensemble délégué. Même chose pour les candidats. Comme Sarmad l'a dit tout à l'heure, c'est en fait impossible à résoudre par le biais d'une révision manuelle. Il y en a trop. Les variantes, vous savez, c'est presque infini, surtout si c'est une longue étiquette. C'est d'ailleurs assez stupéfiant de voir le nombre de variantes qui peuvent être créées. Et cela ne se règle pas simplement en les regardant. Aujourd'hui, on est à 1 600 TLD de zone racine et même ça, c'est trop.

Ensuite, on a la confusion ou le mélange des variantes et des « confusables ». On a ici le premier mot qui se trouvait en français, aéronef, avec une cartographie. On a traduit en système ASCII « aeronef » sans l'accent. Ce n'est pas une variante ici, c'est une série de

caractères similaires ; ainsi, cela se voit au moins dans la zone racine. Et la dernière ligne, c'est là qu'on a une confusion avec le « AE ». Ce n'est pas une variante de A ou de E, mais on a quand même cette similarité visuelle. Ce n'est pas nécessairement « AE », il n'y a pas de « AE » en français par exemple, mais on a le O dans le E et à ce moment-là, on a une similarité entre le O, le E et la paire E dans l'O. Le mot « cœur » est un exemple éclatant en français. L'E dans le O était séparé pendant très longtemps en français. Il y a encore des gens qui se trompent et qui oublient l'E dans l'O. Voilà la source de confusion ici.

Maintenant, il est très important de déterminer le niveau ou les étiquettes de similarités. Il y a cinq niveaux. Tout d'abord, on a identiques ou quasi identiques. C'est en fait le [inaudible] du système des variantes puisqu'on revendique un caractère identique. Ensuite, on a le fort niveau de confusion en numéro deux. C'est assez fréquent, on a vu l'exemple avec le F tout à l'heure. Ensuite le numéro trois, c'est la zone un peu grise et floue, ces étiquettes similaires où la plupart des cas sont similaires. Ensuite, en numéro quatre, similarités distantes. Certaines personnes penseraient qu'ils sont similaires, mais très souvent, ils ne le sont pas. Et le cinquième niveau, étiquettes distinctes – cet exemple le montre à l'écran justement. Le premier, par exemple, on a le latin, le .cap et ensuite cyrillique .cap qui serait .sar en cyrillique. Deuxième exemple, on voit POP, les trois lettres qui ressemblent à POP en grec et ensuite POP en latin. Finalement, on a le troisième exemple, P en latin et RA en cyrillique. Et en dernier, c'est un exemple assez folklorique avec les signes chinois mais qui décrivent les deux mots

« domain label » en anglais. Évidemment, cela ne confondrait personne.

Une autre considération intéressante, il y a certaines lettres simples comme le O, C, L, S, V, X qui sont très partagées entre les scripts. On voit que le cercle qui serait un O est retrouvé dans huit alphabets. Peut-être en hébreu, ça change un peu, mais sinon ça se ressemble beaucoup. Si vous avez un domaine qui a juste O, là, il y a confusion maximale pour beaucoup d'alphabets parce que cette lettre se retrouve un peu partout. Il y a plusieurs combinaisons qui existent. Prenons l'exemple du COCO en latin en bas de la page, SOSO en cyrillique et puis ce qui ressemble à COCO en birman, ils ne ressemblent, ils ne veulent pas dire la même chose du tout, mais visuellement, c'est très similaire. Il faut en tenir compte. Et c'est encore pire quand on répète des lettres comme le O. On a l'exemple dans la racine. Aujourd'hui, on a trois O je pense, et ce O triple bloque beaucoup de variantes dans d'autres scripts. On ne peut pas le négliger. Il y a aussi le fait que beaucoup de variantes latines bloquent le cyrillique et vice versa à cause de la zone racine.

On me demande d'accélérer quelque peu. C'était juste un petit exemple. Je ne vais pas non plus m'appesantir. On voit ici à gauche qu'on a le latin. À droite, on a une lettre grecque et tout en bas, je pense que c'est l'arménien. Ça se prononce G ou GUE. C'est un bon exemple. Cela nous montre l'espace de perception de la situation. On a deux lettres qui semblent un peu différentes à gauche et à droite à cause de la cédille qui part à gauche et à droite, alors que celle d'en bas dans le cercle rouge part directement vers le bas. C'est ce qu'on appelle l'espace perception. Mais c'est un peu un mauvais exemple parce

qu'elles ne sont pas considérées comme des variantes par le système LGR. Peu importe. Prochaine page.

Ensuite, la partie des majuscules et des minuscules. Le navigateur en général fait cela, majuscules et minuscules. Et dans les situations où nous avons des majuscules ou une introduction des majuscules, cela peut donner lieu à des confusions entre le grec et le latin, parce que quand on est en minuscules, il peut y avoir des similarités. Vous voyez cela avec le H. Quand on passe du h minuscule en grec ou H majuscule, on a une similarité avec le grec, la même chose pour le N, on peut passer à un V. Vous voyez qu'on ne peut pas éviter cela. Il y a une situation ici qui va avoir lieu. Je voulais vous montrer un peu ce que cela peut donner quand on accepte les majuscules et les minuscules, il faut le savoir. Il faut accepter aussi le fait qu'il y a des limites concernant la possibilité d'éviter cette confusion entre majuscules et minuscules. Il y a beaucoup trop de faux positifs.

Ensuite, il y a des séquences qui portent à confusion. Ici, vous voyez un bon exemple RN versus M. Cela a été utilisé dans le passé pour des activités d'hameçonnage d'ailleurs. Myspace justement sur Internet a souffert un hameçonnage. Ensuite, on a des séquences dont je vous ai parlé et vous allez voir qu'il y a des caractéristiques répétées lorsque les caractéristiques sont répétées et difficiles à détecter parce qu'elles sont au milieu. Vous avez cette situation dans laquelle en birman vous avez une confusion. Il est difficile de reconnaître un mot de l'autre.

Ensuite, il y a des contextes dans lesquels on analyse la similarité des chaînes et en fonction des scripts. Il faut tenir compte du script aussi.

On part toujours du principe que dans des situations de confusion, l'utilisateur cible et connaît le script. C'est un point important. On parle toujours du principe qu'on n'a pas une prise en compte mondiale de cette confusion ; c'est seulement pour les personnes qui connaissent la langue, mais c'est quand même dangereux, parce qu'il y a un problème cérébral qui fait qu'on ne le voit pas. Prochaine diapositive.

En fonction de l'utilisateur, la similitude est dans l'œil de celui qui regarde. On a fait des études concernant ce qui pourrait être considéré comme visuellement similaire et on est parvenu à cette confusion. Nous avons déterminé que certains scripts sont liés par leur histoire et leurs caractéristiques. C'est le cas du latin, du grec, du cyrillique et de l'arménien. Il y a des formes qui portent à confusion ; la même chose pour le chinois, le japonais, le coréen, le hanja dans le cas du coréen. En réalité, il y a une grande différence entre le chinois et le japonais. La même chose pour les scripts néobrahmi utilisés en Inde avec des origines communes et des lettres qui se ressemblent. Il y a une tentative pour harmoniser ces scripts qui fait que l'on a encore plus de confusion.

Ensuite, il y a des scripts qui se ressemblent, par exemple le kannada et le telugu qui ont des formes qui se ressemblent, le latin et le cyrillique, il y a une question d'accent, une série d'aspects, et le devanagari et le gurmukhi. Il y a certains cas dans lesquels le thaï et le laotien sont aussi étroitement liés, mais les styles de polices typiques sont différents.

Ensuite, nous avons des ressemblances plus faibles. Ici, vous voyez des confusions que nous avons constatées entre han versus katakana et hiragana s'il s'agit de système de variantes dans la zone racine. Si vous

regardez les similarités qui sont visuelles, je ne vais pas trop rentrer dans le détail ici, mais il y a des cas où il n'y a aucune similarité. Il y a certains scripts dans lesquels on n'a pas de problème interscript. C'est le cas de l'arabe, de l'éthiopien, du gujarati, du khmer, du laotien, du thaïlandais.

Voilà quelques exemples. Voyez kannada versus telugu, deux chaînes et on ne voit pas vraiment la différence. L'arménien aussi. Vous voyez ici l'écriture traditionnelle arménienne qui ne ressemble pas au latin, mais quand on regarde la forme moderne de cet arménien, c'est très proche du latin, pas en tant que mot parce que c'est tout à fait différent, mais il y a une ressemblance avec les caractères latins, les U. Il y a beaucoup de mots latins qui pourraient ressembler à des mots en arménien. En grec, la même chose, en fonction du type d'écriture qu'on va utiliser, traditionnel ou non, si l'on utilise le grec moderne, on va avoir des lettres qui vont être semblables. Et la même chose pour l'éthiopien et l'arménien : si vous regardez une lettre toute seule, il peut y avoir des confusions. C'est la même chose pour des ensembles de deux lettres ou de trois lettres, il peut y avoir des confusions.

Voyons un peu maintenant le processus qui va donner lieu à la discussion. On a fait un examen de l'évolutivité de la similarité de chaînes. Il faut mettre en place un processus en deux étapes. On va utiliser quelque chose qui va être basé sur le système de variantes dans la première phase et dans la deuxième phase. Après la vérification, un panel d'experts va analyser les sous-ensembles de niveau un que l'on a pour faire une combinaison finale si cela est nécessaire. Parfois, on n'en

a pas besoin. Parfois on va faire une analyse et on va se rendre compte qu'on peut le différencier.

La même chose pour la taille de l'opérateur de registre. Vous voyez, j'ai fourni quelques chiffres ici concernant les tables IDN et j'ai utilisé une source dans laquelle on voit les scripts. On a 1 590 entrées et la plupart d'entre elles, 1 422 entrées sont ASCII et 168 sont des entrées IDN. Cela montre un peu l'équilibre qu'il y a entre ASCII et IDN. Beaucoup d'ASCII ont de nombreuses variantes ; ce n'est pas toujours très clair, mais elles existent. Par exemple, vous les voyez ici AAA, ARP, AC, ACO qui pourraient être confondus avec des étiquettes en cyrillique. Ensuite, on a d'autres termes cyrilliques comme OPR qui peuvent être confondus.

Ensuite, si vous regardez par script, vous voyez qu'on a plus de différences entre latin, [inaudible] – c'est surtout le chinois –, l'arabe, 36 en arabe. Cela va en diminuant avec le cyrillique, [inaudible], etc. On a une entrée pour le nom en tant que pays. C'est pour cela que l'on a ici cette représentation. Ici, vous voyez la différence entre ASCII et non-ASCII. La taille de la zone racine doit être suffisamment petite pour permettre la mise en œuvre des stratégies décrites ci-dessus.

Les conseils pour utiliser ce système de variantes. Nous recommandons d'utiliser ce que nous appelons l'index, les niveaux de variantes et index. On appelle cela aussi parfois le squelette. On va associer chaque niveau qui a une valeur unique de façon à ne pas être obligé de comparer. On va avoir une valeur qu'on va comparer à deux niveaux. C'est un mécanisme très puissant qui va nous permettre d'éviter de

comparer, parce que comparer demande beaucoup de temps. Ces systèmes de variantes sont très utiles pour cela.

BILL JOURIS : Quand j'ai lu votre document pour préparer cette séance, je n'ai rien vu sur le calcul que vous faites pour calculer cet index. Je n'arrive pas vraiment à comprendre comment il est possible de faire ce calcul.

MICHEL SUIGNARD : Nous utilisons les points de variante. L'index ne doit pas être un niveau valide. Nous calculons cela dans le niveau possible. Vous prenez un ensemble de variantes et vous utilisez le point le plus bas. C'est assez arbitraire. On pourrait faire le contraire, mais en général, on fait comme cela. La plupart du temps, c'est une table qui est valide et si ce n'est pas le cas, nous avons des index.

Nous faisons la même chose pour les prévérifications. Nous créons des grappes et nous les traitons comme des ensembles de variantes. Dans les variantes, nous utilisons quelque chose qui s'appelle la cartographie entre un point, une séquence. C'est un mécanisme qui nous permet de créer une disposition dans la zone racine. On pourrait utiliser cela comme mécanisme pour le triage des différents niveaux pour analyser tout ce qui est basé sur la perception. Notre objectif est d'avoir un ensemble de variantes et de faire une cartographie pour déterminer s'il y a des similarités, si cela est identique, etc. Nous faisons cela et ensuite, nous mesurons les résultats. La mauvaise nouvelle, c'est qu'il y aura peut-être des confusions ici.

J'avance alors, prochaine page. Quel genre de données avons-nous ? Beaucoup de données sont venues avec les études de zone racine on a des groupes comme LGCA qui ont donné beaucoup de données. On n'a pas grand-chose sur les séquences, c'est un domaine qui demeure inexploré. Bien entendu, le résultat de la révision de similarité de chaîne n'est pas éligible parce qu'évidemment c'était une variante, ce n'est parfois même pas valide. Pour le passage, ce n'était pas problématique du tout. Aucun problème à noter. Ensuite pour ce qui est du litige, c'est une situation où on a deux personnes qui ont fait une demande. Et ensuite, pour l'échec, c'est quand on a une confusion avec une étiquette réservée ou existante.

Euristique de vérification. Encore une fois, on a une échelle de 1 à 5. Il faut s'appuyer sur le système LGR pour voir comment on s'y prend, mais c'est le même système que pour la zone racine. Pour comparer, on utilise l'action pour déterminer leur niveau de proximité dans l'espace de perception. Pour que cela marche, il faut faire comme avec la zone racine LGR, il faut créer des LGR, donc il faut prendre les 26 LGR qu'on a avec la zone racine et utiliser une autre situation de cartographie, une autre méthodologie en utilisant la même structure, le même répertoire auquel on applique les variantes en utilisant un nouvel ensemble de LGR qui seront internes. Ils sont alternes au mécanisme et c'est comme cela qu'on crée un système de grappes. Comme j'ai dit avant, on utilise un indice de grappes par appartenance et ensuite, on analyse cette grappe.

Cela fait beaucoup de textes. Je vais essayer de résumer. Voici un autre système utilisable. Ce pourrait être utilisé dans certains contextes. Au

lieu d'utiliser le système de variantes, on utilise des étapes existantes. Vous verrez les détails dans le document. C'est en fait un autre moyen d'évaluer la similarité en utilisant des étapes de modification. On peut enlever l'ottoman par exemple, on enlève une lettre ou on enlève des lettres doubles. C'est un autre système qu'on pourrait envisager pour la prévérification des similarités. Mais je dois accélérer.

On voit la postvérification, c'est une évaluation manuelle. Ici, nous allons utiliser beaucoup d'informations créées par la prévérification pour déterminer tout d'abord que le nombre que vous évaluez est assez bas et que vous validez la prévérification. Ensuite, vous pourrez injecter dans votre propre étiquette pour la révision. Parfois, certains diront : « Vous auriez dû envisager ce niveau parce que ce n'est pas ambigu pour la prévérification », mais cela peut porter à confusion dans d'autres contextes.

Ici, on a une marge de manœuvre pour ce qui est d'ajouter des chaînes. C'est un peu comme quand on passe de 2 000 à 1 000 à 50 ou 60, c'est beaucoup plus facile pour le panel. Et le panel n'utilisera pas de groupe d'experts mais des experts individuels en scripts ou en alphabets. Il faut au moins trois examinateurs dans le panel pour la postvérification, il faut qu'ils connaissent ces scripts. Il faut qu'ils connaissent le DNS. Ils doivent être capables de lire un score informatisé et de le contester si c'est nécessaire. Et pour les résultats, cela peut être rejeté si c'est trop facilement confondu avec des étiquettes préexistantes ou existantes.

Les prochaines étapes consistent à affirmer ce qu'on fait et de publier des commentaires publics. Nous allons faire peut-être aussi une phase

d'essai avec un ensemble de 1 500. On applique aussi un examen ou une révision de similarités pour voir s'il y a des problèmes de similarité et ensuite, évidemment, on avance. La dernière page, je pense que c'est juste les liens. Merci.

SARMAD HUSSAIN : Des questions maintenant. Pitinan s'occupera de la liste des questions.

PITINAN KOOARMORNPATANA : Oui, n'hésitez pas à lever la main si vous avez des commentaires ou des questions. Première question sur Zoom.

ABDALMONEN GALILA : Attention, sur mon Zoom, il n'y a pas de main levée.

PITINAN KOOARMORNPATANA : Oui, j'ai quelqu'un qui surveille dans la salle aussi. Bill Jouris.

BILL JOURIS : J'aime bien les cinq niveaux de similarités que vous avez ajoutés. Ce que je ne sais pas trop, c'est comment on peut attribuer ces niveaux. Quand le panel de production d'étiquette décide si les choses étaient des variantes ou pas, cette étape nous a pris cinq ans. Ici, j'ai l'impression que c'est le même effort colossal. Comment on va réussir à faire cela avant le début de la prochaine série ?

MICHEL SUIGNARD : C'est une bonne chose que vous y ayez consacré cinq ans comme ça, on sait apprendre de ses erreurs. Je pense que le LGR latin a fait un travail monumental là-dessus. Nous, évidemment, on va s'appuyer sur ce travail déjà fait. Certains panels de générations ont fait ce travail et nous allons en tirer parti, aucun doute. Bien sûr, on n'essaie pas de réinventer la roue. Bien entendu, il y a des décisions arbitraires ici. Parfois, il faut trancher. Moi, j'y participe beaucoup. Et nous sommes au courant. On s'y connaît mieux parfois dans certains alphabets latin, grec, cyrillique ; on connaît la chanson plutôt bien même pour CJK, je m'y connais plutôt bien. Je peux voir la similarité visuelle dans les chaînes CJK. Parfois, on se tourne vers d'autres experts. Là où on pêche beaucoup, c'est le néobrahmi pour l'Inde. Et en arabe, notre groupe d'experts va nous aider parce que l'arabe est un sujet assez complexe et je ne prétends pas avoir des connaissances en la matière. Je m'attends à ce que d'autres groupes s'en chargent.

PITINAN KOOARMORNPATANA : Alan.

ABDALMONEN GALILA : Deux questions. Tout d'abord, les sections 9.1 et 9.2 étaient un peu floues. La taille de l'opérateur de registre, ce n'était pas tout à fait clair. Est-ce que vous avez une explication ? Parce que j'ai une deuxième question aussi.

Ensuite, étant donné qu'il s'agit de similarités visuelles, on a enregistré beaucoup de progrès dans tout ce qui est reconnaissance de schémas

et de patterns, comme on dit en bonne informatique. Je m'attendais à ce qu'on en parle ici, même si c'est écarté éventuellement.

SARMAD HUSSAIN : Deuxième question, j'y réponds. Je pense que c'est une excellente recommandation. D'ailleurs, on recommande de ne pas utiliser des outils automatisés. Bien entendu, il faut que le processus soit manuel. Merci.

ABDALMONEN GALILA : Oui, je sais pourquoi et je comprends pourquoi on n'a pas utilisé le dernier. Mais le monde a beaucoup changé depuis ; vos téléphones reconnaissent vos visages maintenant, ce n'était pas possible il y a 10 ans.

MICHEL SUIGNARD : Nous reconnaissons ces récurrences, ces schémas. Il en est question dans les lignes directrices également. Pour ce qui est de la taille du registre, on ne regarde que la zone racine, pas le deuxième niveau. C'est la liste de la zone racine. Je ne sais pas ce que vous voulez dire par deuxième niveau.

ABDALMONEN GALILA : Les sections 9.1 et 9.2 parlent de l'importance pour un grand registre. Il est dit que ce serait complexe. Moi, je ne comprends pas pourquoi on en parle.

MICHEL SUIGNARD : Parce que quelqu'un pourrait s'en servir. Vous pourriez essayer d'évoluer vers un plus grand registre. Mais pour le moment, ce n'est pas la portée de cette application.

PITINAN KOOARMORNPATANA : Merci.

Nabil, allez-y.

NABIL BENAMAR : Petite question pour l'alphabet arabe. Moi, je fais partie de l'équipe de groupes d'experts pour les IDN quant aux lettres arabes. On a fait partie du travail qui a été mis en œuvre pour la zone racine pour les domaines de premier niveau et aussi pour le deuxième niveau pour plusieurs langues. Je me demande ce qui manque pour que nous puissions donner un nouveau souffle à cet effort de groupe d'experts.

MICHEL SUIGNARD : La portée de ce travail diffère un peu de sa version antérieure. Avant, on regardait les variantes, c'est-à-dire la première catégorie des cinq. Maintenant, nous regardons les trois premières catégories de ces cinq. Étant donné que la portée change, peut-être qu'après analyse, le panel de génération ou la communauté des scripts estimera que rien ne doit être ajouté ou identifié pour une révision de similarités. Mais encore une fois, il faudrait peut-être faire un dernier passage pour veiller à ce qu'il n'y ait pas d'autres points de code supplémentaires à identifier.

PITINAN KOOARMORNPATANA : Merci.

Allez-y, Bill.

BILL JOURIS : Tout d'abord, je suis très content de lire ce document que vous avez mis sur pied. On est vraiment très au courant maintenant. J'avais plusieurs griefs et j'ai trouvé presque toutes mes réponses et mon soulagement dans ce document point par point, et c'est très bien.

Toutefois, je pense que vous avez omis une chose. Tous les grands navigateurs, Firefox, Chrome, Bing, quand vous avez un nom de domaine, ils changent la couleur, ce n'est pas un problème et ils vont le souligner. Et les grands traitements de texte comme Microsoft Word et Acrobat font la même chose. Les symboles qui pourraient bloquer, par exemple dans l'alphabet latin, on a le O souligné. Si vous soulignez un nom de domaine qui en comprend un, il y aura des petits points de pixel de chaque côté de cette ligne. Mais comme ce caractère n'apparaît que dans certaines langues de petite diffusion, l'utilisateur va voir cela et ne va même pas savoir qu'il y a quelque chose à voir, qu'il y a une interruption dans cette ligne. Cette sous-ligne est très importante.

MICHEL SUIGNARD : Oui, nous en savons quelque chose parce que quand nous faisons la zone racine, nous connaissons ce problème. Ce que vous dites, c'est la même chose pour le latin, pour le cyrillique, en Europe en général. Je

pense que c'est tout à fait vrai et cela dépend du mandat aussi. Nous ne sommes pas chargés non plus de s'occuper des mots soulignés. Mais si la communauté y trouve un intérêt particulier, on va s'y pencher, c'est sûr.

SARMAD HUSSAIN : Oui, tout à fait, bonne suggestion. Et comme Michel l'a bien dit, nous attendons les indications de la communauté. Si vous pensez qu'il y a des choses à ajouter, s'il vous plaît, formulez un commentaire public et cela ne passera pas inaperçu.

BILL JOURIS : Oui, c'est un travail en cours.

MICHEL SUIGNARD : On insiste souvent sur les singuliers et les pluriels, même si c'est prometteur. Mais c'est encore très anglophone comme vision du pluriel, même si ce n'est pas toujours vrai en anglais. Quand ils utilisent des mots français, ils pluralisent ces mots différemment. Le pluriel comme concept, c'est vraiment un terrain dangereux. Je ne sais pas si on devrait s'y aventurer. À l'heure actuelle, on ne préfère pas y toucher.

PITINAN KOOARMORNPATANA : Merci Michel. Nous sommes arrivés à la fin de l'heure impartie, mais nous prenons la dernière question.

BEN MCELWAINE : Ben du registre Google. Visiblement, il s'agit juste de la similarité visuelle des scripts. J'aimerais savoir si quelqu'un travaille sur ce qui est en dehors de la portée. Il a d'autres sources de confusion qui sont encore pires parce qu'elles s'appuient seulement sur les caractères latins et rien d'autre. Par exemple, la pluralisation, si on a par exemple le mot pièce de monnaie et pièces avec un s, et bien vous pouvez mal entendre une adresse e-mail. Vous savez, dans beaucoup de langues du monde, il n'y a pas de pluriel tout simplement, le concept n'existe pas. Les gens qui ne parlent pas anglais comme langue maternelle pourraient rater cette nuance pluriel-singulier. Ils pourraient aller sur les sites Web avec le mot au singulier et s'égarer complètement.

Ensuite, on a aussi les variantes d'un pays à l'autre. Les Américains écrivent le mot couleur sans le U alors que les Britanniques mettent un U. Ce serait très bizarre d'avoir un .color sans Y, *color* et *colour* qui est le même mot en anglais mais écrit différemment, et qui te dirige vers deux sites différents. Il faut tenir compte de tout cela. Est-ce que les gens préfèrent que ce soit délégué ou est-ce que les gens s'y opposent ? Qu'en pensez-vous ?

SARMAD HUSSAIN : Merci Ben pour cette question. Il y a une recommandation en la matière qui concerne le singulier et le pluriel et c'est évalué en ce moment par la communauté, par le conseil d'administration. Et à part cela, il n'y a pas d'autres recommandations sur une portée élargie de la similarité qui pourrait, comme vous l'avez dit, dépasser le simple visuel de la similarité. Par conséquent, tout dépend des types de similarités. Il y a

aussi une question liée aux numéros. On peut avoir aussi des variations en genre, en morphologie.

En ce qui concerne notre travail, de toute façon, je dirais que nous sommes basés sur les recommandations qui nous ont été fournies. Nous travaillons sur ces recommandations au niveau politique.

MICHEL SUIGNARD : Nous avons mentionné certains exemples de similarités, mais on pourrait en mentionner d'autres. Il y a tellement de mots en français et en anglais qui ont des orthographes différentes avec des consonnes simple ou double par exemple. Vous avez parlé du O ou et du OU, mais il y a énormément d'exemples de confusion possible en fonction de l'orthographe. Il nous faut voir quelle est la dimension sur laquelle nous voulons travailler. Il est clair que le panel, à la fin, peut ajouter certains aspects manuellement dans ce processus. Vous avez parlé de ces confusions entre *color* ou *colour*. Ce peut être introduit de manière manuelle aussi, si cela est nécessaire. Notre prévérification vise à étudier la question pour qu'on puisse ensuite commencer à travailler au niveau manuel pour ce qui ne peut pas être corrigé.

SARMAD HUSSAIN : Nous allons maintenant conclure cette séance. Je vous remercie tous pour votre participation. Merci Michel pour votre présentation. Si vous avez des questions à nous poser, n'hésitez pas à nous joindre à notre adresse de icann.org. N'hésitez pas non plus à participer à la période de

commentaires publics. Merci. Cette réunion est maintenant terminée et nous pouvons arrêter l'enregistrement.

[FIN DE LA TRANSCRIPTION]