ICANN82 | Forum de la communauté – Évaluation de la similarité des chaînes et prochaine série du programme des nouveaux gTLD
Mercredi 12 mars 2025 – 09h00 à 10h00 PST

PITINAN KOOARMORNPATANA Très bien. Donc, nous allons commencer. Nous allons lancer l'enregistrement. Nous avons déjà lancé l'enregistrement. Allez-y. L'enregistrement a débuté. Merci beaucoup.

Bienvenue à cette séance d'évaluation de la similarité des chaînes dans le nouveau programme gTLD. Je m'appelle Pitinan, je suis participante et je gère la responsabilité à distance. Cette séance va être enregistrée et les règles de comportement de l'ICANN seront prises en compte durant cette séance.

Les questions et les commentaires ne seront lus que si [elles] sont dans la fenêtre questions-réponses. Si vous le mettez dans le tchat, cela fera partie du tchat et nous pourrons éventuellement lire cela à haute voix. Mais cela n'est pas sûr. Nous aurons l'interprétation en anglais, français et espagnol. Et si vous voulez prendre la parole durant cette séance, veuillez lever la main sur Zoom. Nous vous donnerons à ce moment-là la permission d'ouvrir votre micro. Dans la salle, utilisez les micros, s'il vous plaît. Indiquez votre nom pour l'enregistrement, la langue dans laquelle vous allez vous exprimer s'il ne s'agit pas de l'anglais, et parlez à un rythme raisonnable.

J’aimerais maintenant passer la parole à M. Sarmad Hussain.

SARMAD HUSSAIN Oui, merci beaucoup. Pitinan. Bienvenue à toutes et à tous. Bonjour ou bonsoir où que vous soyez sur le globe.

Donc, pour vous donner un aperçu de notre ordre du jour, je vais commencer par présenter très brièvement un aperçu sur la similarité des chaînes et sur la portée de ce programme d’évaluation de similarité des chaînes. Et ensuite, nous allons demander à Michael Bauland de nous aider en ce qui concerne les efforts qui existent pour collecter des données sur les similarités de chaînes.

Donc, nous aurons une mise à jour à ce niveau. Et Julian Bernard prendra la parole, ainsi que David. Ils parleront donc d’outils d’évaluation, d’outils qui sont donc mis en œuvre et conçus maintenant pour évaluer la similarité des chaînes. Et ensuite, nous parlerons du calendrier que nous aurons. Vous aurez la possibilité de poser des questions par la suite, mais n’attendez pas la fin de la présentation pour poser des questions. Si vous avez des questions durant les présentations, n’hésitez pas à lever la main ou de mettre une question sur le cadre questions-réponses de Zoom. Vous pouvez ici également lever la main.

Donc nous allons sans plus attendre commencer notre présentation et c’est donc un aperçu que j’aimerais vous donner concernant la prochaine série de gTLD. Nous allons avoir beaucoup

de travail à effectuer sur ces différentes chaînes pour nous assurer que les chaînes que l'on demande donc ne sont pas similaires à des chaînes qui existent déjà. Et nous devons travailler également aux confusions possibles. Et nous voulons nous assurer que les utilisateurs finaux ne soient pas en pleine confusion par rapport à ces similarités de chaînes qui pourraient exister. Donc c'est un aspect visuel en fait.

Donc la similarité peut être également au niveau du contenu, du sens, mais nous nous sommes limités à l'aspect visuel de la similarité des chaînes. Nous avons quelques affirmations dans les procédures ultérieures. SubPro.

On nous a donc demandé que cette similarité des chaînes soit faite comme ça a été le cas en 2014, que cela soit fait pour la prochaine série de gTLD. Donc nous avons plus de détails sur la portée de tout ce travail dans le rapport final de la première phase [des] PDP, donc, sur les noms de domaine internationalisés. Et je vous encourage à lire ce rapport final, à voir les recommandations spécifiques. Il y a une vraiment -- un résumé uniquement sur cet écran et regardez le libellé de ces recommandations dans les différentes politiques et dans les différents documents et rapports finaux.

Donc, pour vous donner un simple aperçu aujourd'hui, avant de rentrer dans les détails, la portée des comparaisons, il y a des différences entre les variantes et ce que j'appelle les chaînes similaires. Donc les variantes sont définies dans les règles de

procédure comme des chaînes qui sont considérées comme étant les mêmes par une communauté spécifique qui utilise ces écritures, ces scripts et on ne peut pas distinguer donc ces deux chaînes. C'est ce qu'indique cette communauté linguistique. Donc il y a des homographes, par exemple, qui sont similaires visuellement, mais également au niveau du sens parfois. Donc -- donc, ça, c'est les règles de génération d'étiquette pour la zone racine. Et cela est déjà défini comme variante.

Et il y a des chaînes supplémentaires qui pourraient être similaires sans être exactement identiques et qui sont définies comme des variantes. Mais cela peut être donc -- ça peut prêter à confusion pour les utilisateurs finaux.

Donc, en ce qui concerne l'évaluation de la similarité des chaînes, eh bien, nous voyons ces ensembles de chaînes qui ne sont pas des variantes, mais qui prêtent à confusion. Donc voilà donc la portée que nous avons dans notre travail. Donc, sur la base des recommandations qui sont dans le PDP des IDN, vous avez —— en haut sur l'écran, vous avez les chaînes qui sont demandées. Cela peut être des gTLD IDN primaires, [qu'elles] soient en ASCII ou IDN, qu'elles soient en caractères ASCII ou noms de domaine internationaux. Il y a également donc toutes les variantes en bloc. Donc vous avez trois ensembles, donc les chaînes primaires gTLD, toutes les variantes qui peuvent être allouées et toutes les chaînes variantes bloquées.

Donc la question, c'est la comparaison de ces chaînes. Donc vous voyez, sur la gauche, vous avez les catégories de comparaison pour déterminer les similarités. Donc, ceci inclut les gTLD existants, la chaîne gTLD qu'on avait déjà demandée lors d'une série précédente, les ccTLD [existantes], les IDN requis au niveau des ccTLD de noms bloqués, et il y a des sous-catégories de noms bloqués qui s'appliquent. Et il y a également les caractères ASCII. Et quand il y a deux caractères ASCII uniquement, ça, c'est pour les IDN, c'est pour les noms de domaine internationalisés au niveau des ccTLD. Donc toutes ces chaînes vont être comparées par rapport aux chaînes primaires et par rapport, également, à -- toutes les variantes pourront être allouées. Et cela va permettre d'effectuer des comparaisons entre les chaînes demandées et toutes les variantes.

La seule comparaison qui ne rentre pas dans le cadre. C'est en bas à droite. C'est en noir. Les comparaisons entre les chaînes de variantes bloquées et toutes les chaînes de variantes bloquées, donc au niveau des chaînes pour lesquelles on a fait une demande de gTLD. Donc, vous voyez, lorsque c'est en gris, ça, ce sont des comparaisons. Avec le panel ou le panel qui travaille sur la similarité des chaînes, donc on ne va pas penser que cela soit pertinent, que cette comparaison soit pertinente au niveau, en tout cas, de la confusion des chaînes.

Donc, vous voyez tout ce qui rentre donc en ligne de compte, tout ce qui fait partie de ce travail. Nous avons publié des directives initiales pour conduire une révision de la similarité des chaînes qui

[a été publiée] pour commentaires publics, il y a de cela quelques mois. Nous nous sommes basés sur le travail qui se poursuit et qui avance. Et nous allons avoir une nouvelle version de ces directives pour commentaires publics plus tard cette année. Donc, suivez cela de près. Ces lignes de conduite et directives, c'est pour le panel, pour aider le panel dans son travail d'évaluation des chaînes.

Et comme vous l'avez vu, il y a encore des comparaisons qui doivent être effectuées. Il y a la question des ccTLD et des gTLD. Il y a les questions des étiquettes de variantes et des comparaisons entre tout cela. Donc il y a encore beaucoup de comparaisons à effectuer, beaucoup de travail à faire. Donc ce panel d'évaluation des chaînes doit gérer ces comparaisons et doit développer également un outil de préanalyse avec des collectes de données similaires provenant d'experts. Et nous allons parler avec ces experts des scripts et des écritures. Et nous allons utiliser ces données pour créer une estimation sur ce qui peut être considéré comme similaire ou pas similaire. Et ensuite, ces données et cette préanalyse vont être utilisées par le panel d'évaluation pour réeffectuer l'analyse, le contrôle et l'analyse, et prendre une décision finale. Donc tout cela aide ce panel d'experts qui peut statuer et prendre les décisions finales. Et c'est le panel qui va donc déterminer au niveau des chaînes s'il y a similarité ou pas.

Donc on va reparler de cela plus en détail pour divers scripts. Donc là, je voulais simplement présenter les choses. Je ne sais pas s'il y a des questions pour le moment. Sinon, nous allons nous mettre à

parler des données et des outils qui sont utilisés pour analyser ces chaînes et pour évaluer la similarité éventuelle entre ces chaînes.

PITINAN KOOARMORNPATANA Nous avons une question dans la salle.

INTERVENANT NON IDENTIFIÉ Oui, merci beaucoup Pitinan. Merci beaucoup de vos définitions très claires concernant la similarité des chaînes. Merci, Sarmad, pour cet aperçu. Donc vous avez parlé de comparaison. Vous avez parlé de variantes. Donc, est-ce que nous avons. Un outil informatique pour effectuer cette comparaison ou est-ce que c'est uniquement visuellement, manuellement que cela se fait ? Ou est-ce que ça peut être automatisé au niveau informatique ?

SARMAD HUSSAIN Oui, nous sommes en train de développer cet outil d'évaluation de similarité des chaînes pour faire une préanalyse. C'est automatisé. C'est informatisé, et ça fait une liste courte qui ensuite est analysée par les experts. Donc c'est une portée très large. C'est pour ça que nous avons besoin de développer cet outil pour évaluer la similarité des chaînes. Mais les décisions ne seront prises qu'au niveau humain par les experts.

Oui, je crois qu'on a Crystal qui va prendre—

CRYSTAL ONDO Oui, oui. Est-ce que ça a un impact sur le pluriel -- le singulier et le pluriel ?

SARMAD HUSSAIN Donc ça, c'est différent. C'est différent des similarités. Singulier-pluriel, c'est un processus séparé. Le singulier-pluriel, la portée de cela, c'est uniquement l'aspect visuel de ces chaînes -- similarités visuelles.

CRYSTAL ONDO Donc ça va être un débat séparé, le singulier-pluriel.

SARMAD HUSSAIN Oui, tout à fait. Nous avons Abdalmonem. Oui, nous avons quelques mains qui se lèvent. À qui le tour dans la salle ? Abdalmonem. Et les personnes suivantes ? Oui.

TAPANI TARVAINEN Pardonnez-moi. Donc, c'est à mon tour. Est-ce mon tour ?

SARMAD HUSSAIN Non, non, pas encore. C'est d'abord Abdalmonem.

ABDALMONEM GALILA Merci beaucoup pour votre présentation. Alors ma question est une question de suivi par rapport à ce que Dr Anil [Kushi] a dit. Cette évaluation, elle se fera déjà pendant la phase de présélection, de

	préanalyse. Si un candidat passe cette évaluation et un autre candidat tente de faire une demande pour une chaîne qui pourrait être similaire à une chaîne déjà prise, alors ce candidat ne va pas passer par le système de présélection. C'est cela ?
SARMAD HUSSAIN	En fait, l'outil de présélection est utilisé après que les candidatures aient été soumises, et non pas avant. Et cette présélection ne concerne pas les candidats. C'est un outil qui sera utilisé par le panel d'experts pour mener l'évaluation de la similarité des chaînes. Une fois que les candidatures seront soumises, elles vont être passées en revue dans le cadre du processus d'évaluation. Et l'une des étapes de ce processus sera la révision de la similarité des chaînes par le panel d'experts qui utilisera cet outil de présélection et de préanalyse pour déterminer ce qui pourrait être des chaînes qui prêteraient à confusion, pour établir donc des ensembles de similaires prêtant à confusion, qui seront alors placés dans des ensembles conflictuels.
ABDALMONEM GALILA	Donc, ce qui prête à confusion, c'est ce mot de présélection, de préanalyse. Bon, peut-être qu'on pourrait utiliser un autre terme, mais ce terme préscreening — présélection, préanalyse, voilà, cela concerne vraiment [que] l'évaluation qui sera menée par les experts. Mais nous allons peut-être utiliser un autre terme la prochaine fois.

TAPANI TARVEINEN

Tapani Tarveinen du CSG. Alors, pardonnez-moi pour cette question. Car je n'ai pas beaucoup suivi vos discussions à ce sujet. Peut-être que ma question n'est pas pertinente, mais s'agissant de la similarité visuelle, je ne pense pas que -- je pense que peut-être des interlocuteurs natifs de langues différentes vont confondre -- pourraient confondre des chaînes qui leur sembleraient similaires, alors que d'autres ne les confondraient pas. Donc cette similarité, elle est visuelle. Et, moi, par exemple, je sais qu'il y a des caractères qui me semblent similaires et que je ne peux, en fait, les distinguer que si vraiment j'y réfléchis.

SARMAD HUSSAIN

Oui, alors nous avons publié des lignes directrices à ce sujet. Je pense qu'il y a des exemptions qui sont mentionnées dans ces lignes directrices. L'une d'entre elles, par exemple, concerne le fait que la similarité -- que la décision quant à cette similarité sera faite par rapport à un interlocuteur natif ou quelqu'un qui parle bien la langue. Donc si, moi, par exemple, je ne connais pas le chinois, je ne vais pas pouvoir évaluer un script, une écriture. Et s'il y a des personnes qui connaissent l'écriture, mais qui l'utilisent pour écrire en langues différentes, et qu'il y a des variantes possibles, s'il est possible que deux personnes puissent percevoir deux caractères de la même façon ou d'une façon différente parce qu'ils utilisent cette écriture pour écrire en langues différentes, eh bien,

ce sont des points qui seront pris en compte par le panel d'experts. J'imagine que cela entrera dans leur mandat.

Mais vraiment, en fait, je pense que l'évaluation doit se faire sur la base des caractères visuels de la chaîne elle-même. Si d'un point de vue visuel, tout est clair, même si certains caractères ne sont pas utilisés dans une langue bien spécifique, mais que quelqu'un qui connaît cette écriture peut identifier une différence, eh bien, c'est un cas qui sera pris en considération par le panel d'experts.

Aussi, la réponse à cette question est la suivante. Oui, il peut y avoir des différences liées à la langue utilisée, mais nous allons faire en sorte que cet outil se concentre notamment sur la similarité visuelle des chaînes. Et ensuite, ce sera au panel d'experts de juger de cette similarité.

Nous avons ensuite Nabil, et puis d'autres personnes qui souhaitent poser des questions. Alors, nous voulons aussi passer en revue la fin de la présentation. Donc, si vous nous le permettez, nous allons donner la parole à Nabil, puis nous passerons les diapos suivantes en revue, et puis nous reviendrons à vos questions. Donc je vais peut-être cesser d'accepter des questions, et on reviendra aux questions par la suite.

NABIL BENAMAR Merci. Je voudrais ajouter quelque chose concernant la similarité ou la confusion. Cela va bien au-delà du concept des variantes, car, comme on l'a défini pour les variantes, on a la possibilité de faire la

distinction entre deux lettres ou deux étiquettes. Mais pour d'autres confusions potentielles, cela peut être dû au fait que deux lettres sont jointes et qu'elles deviennent alors très similaires à une autre lettre ou à un autre caractère. Donc c'est par exemple un point qu'on a ajouté à la définition de la confusion possible. Et bien sûr, encore une fois, tout n'est pertinent que pour les communautés qui utilisent ces écritures, ces scripts. Donc cela dépend du script utilisé.

SARMAD HUSSAIN Merci. Oui, pardonnez-moi. Vous avez levé votre main, mais peut-être pourrions-nous passer à la suite.

STEVE DELBIANCO Oui. Alors est-ce que le panel d'experts sera autorisé à évaluer la similarité, et ce, même si, par exemple, un élément n'a pas été signalé par l'outil lors de la préanalyse ?

SARMAD HUSSAIN Oui, effectivement. Comme je l'ai dit, c'est un outil d'aide et le panel d'experts pour l'évaluation de la similarité des chaînes. Or, la liberté d'ignorer les suggestions faites par cet outil ou d'inclure dans leurs analyses ce qu'il n'a pas signalé.

STEVE DELBIANCO D'accord. Donc ce n'est pas restrictif. Crystal a parlé des singuliers et des pluriels. Cet outil ne se concentrera que sur les similarités

visuelles. N'est-ce pas ? J'ai bien compris ? Mais pourtant, pour des TLD existants et pour les chaînes qui font l'objet d'une candidature avec un « s » à la fin, est-ce que cela ne pourrait pas être -- est-ce que cela ne pourrait pas prêter à confusion d'un point de vue visuel ? Ou est-ce qu'il faudrait qu'un autre département s'occupe de la question du singulier au pluriel ? Mais est-ce que l'addition d'une lettre en général peut prêter à confusion ? Voilà ma question.

SARMAD HUSSAIN Eh bien, il n'y a pas une réponse à cette question. Par exemple, pourquoi ? Eh bien, parce que, en anglais, on peut ajouter une lettre et créer un nouveau mot « a » auquel on ajouterait un « s ». Le mot deviendrait « as », et alors la signification du mot est différente. Donc est-ce qu'ajouter un « s » à la fin d'un mot pourrait prêter à confusion ? Je pense que c'est le panel qui devra en juger. Cela dépendra aussi de la façon dont cette lettre est ajoutée. Il faudra tenir compte de l'écriture du script utilisé, car, comme on l'a dit, la langue utilisée peut avoir une influence.

STEVE DELBIANCO Oui, mais alors ma question est de savoir, si l'on ajoute un « s » à une chaîne, est-ce que le panel d'experts pourra décider ou conclure que cela prête à confusion?

SARMAD HUSSAIN Oui, si le panel d'experts considère que quelque chose prête à confusion visuellement, il est dans leur mandat de pouvoir dire que

c’est le cas. Mais cela ne concernera pas forcément les singuliers et les pluriels.

Bien, alors je pense que nous allons donc passer à la suite. Encore une autre question. Oui.

ASHLEY ROBERTS S’agissant de l’outil d’évaluation, pour clarifier la question, je pense que vous tentez de faire en sorte que les candidats puissent utiliser cet outil également.

SARMAD HUSSAIN Non, pas encore. Mais nous allons passer en revue cet outil et puis nous reviendrons à votre question. Étant donné que l’outil n’est pas encore totalement développé, nous ne pouvons pas répondre à votre question. Mais lorsque ce sera le cas, nous envisagerons la possibilité de le rendre accessible de manière open source. Mais c’est peu probable que nous le rendions accessible aux candidats en ligne. Cet outil est vraiment conçu pour que le panel d’experts l’utilise. Mais peut-être -- voilà, peut-être qu’on pourra le rendre accessible aux personnes qui souhaiteraient l’utiliser.

Avançons. Je vais maintenant demander à Michael Bauland de passer en revue la question de collecte de données de cette évaluation de la similarité des chaînes.

MICHAEL BAULAND

Merci beaucoup. Je m'appelle Michael et, comme Sarmad l'a dit, je gère actuellement l'aspect collecte de données de notre travail. Les données sont définies par les règles de génération d'étiquette pour la zone racine, car tout ce qui se trouve dans notre répertoire est autorisé pour les TLD. Et ces répertoires — tous nos répertoires — rassemblent 25 scripts, bientôt 26. Car l'écriture, le script thaana vient d'être ajouté au RZ LGR. Et tous ces répertoires doivent être analysés.

Mais bien sûr, c'est bien trop de données pour qu'une seule personne ou une seule équipe ne s'occupe de l'analyse de ces données. Par ailleurs, cela exigerait que l'équipe dispose et des connaissances -- connaisse toutes les écritures qui sont analysées aussi. Cette tâche sera déléguée à des experts pour chaque script. Chacun et chacune de ces experts et expertes disposera d'un mandat. Et ils devront comparer leur répertoire aux caractères ASCII, donc les lettres de A à Z, qu'il s'agisse de minuscules ou de majuscules.

Ensuite, il faudra bien sûr aussi effectuer des comparaisons dans un même script. Il y aura aussi potentiellement des comparaisons au travers de différents scripts, car certaines écritures appartiennent à la même famille. Il y a donc des similarités entre ces différents scripts, et les experts devront donc comparer leurs propres répertoires à d'autres répertoires de langues de la même famille. Et ensuite, on se penchera sur les cas divers. Par exemple, on comparera les éléments en majuscules au sein d'un même script — si ce script permet l'utilisation de majuscules, les

conjonctions des caractères composés… Les formes simples devront également être comparées. Par exemple, des cercles, des lignes, des courbes. Il faudra aussi se pencher sur les versions soulignées. Et cela a été inclus dans notre analyse. Car si on se penche sur les noms de domaine, très souvent, il y a un soulignement. Et ce soulignement peut -- donc le soulignement d'un hypertexte d'un hyperlien peut dissimuler le fait qu'un caractère est souligné ou non. Alors que si ce n'était pas souligné, eh bien, il serait simple de les comparer.

Donc on nous a demandé également d'effectuer des catégories, de classer en catégories, donc en cinq catégories, que nous allons donc définir en ce qui concerne les similarités, lorsque c'est identique ou presque identique, donc homographes, et également toutes les variantes déjà définies, donc au niveau des règles de génération d'étiquette pour la zone racine. Deuxièmement, ce qui est qu'on peut confondre fortement avec une forte similitude, presque une homographie. Troisièmement, similaire 4 et 5, c'est peu similaire et distinct.

Cinq. C'est -- pour la majorité des experts, c'est totalement distinct. Ce n'est pas similaire. Et au niveau de la quatrième catégorie, il n'y a que peu de similarités qui existent. Et on peut donc clairement le voir. Et il a été décidé que cela ne faisait pas partie des trois premières catégories.

Donc je vais maintenant rentrer un petit peu plus dans les détails concernant ces catégories et vous fournir quelques exemples,

également, qui ont été créés par des experts de ces scripts ou écritures. Donc je vais essayer d'utiliser certaines écritures qui sont assez diverses, et je vais donc vous montrer d'ici peu ces exemples par rapport aux données qui ont été collectées.

Donc première catégorie, donc identique ou presque identique, plus les variantes. En ce qui concerne le premier point, toutes les variantes existent dans cette catégorie. Les variantes déjà définies automatiquement appartiennent à cette catégorie, indépendamment de leur apparence visuelle. Donc nous ne prenons pas en compte la capitalisation des termes à ce niveau presque identique, donc, où l'on peut facilement les confondre en ayant les deux chaînes l'une à côté de l'autre.

Diapo suivante, s'il vous plaît. Donc pour cette catégorie, un des exemples qui proviennent de l'hébreu, du latin et du bengali. Je ne vais pas les lire parce que je ne pourrai pas les prononcer. Mais vous voyez que, ces chaînes, on peut facilement, d'un côté, les distinguer. Mais imaginez que vous avez par exemple simplement un « i » qui est souligné dans le caractère latin et donc ce n'est pas exactement la même chose.

Nous allons passer à la diapo suivante. Là aussi, dans cette deuxième catégorie, on peut fortement confondre. Une majorité d'utilisateurs qui sont attentifs lorsqu'ils ne regardent pas les deux chaînes l'une à côté de l'autre, eh bien, vont les confondre. Et vous avez deux exemples qui proviennent de l'éthiopien et du latin.

Voilà. Je laisse ça un instant à l'écran pour que vous le voyiez visuellement. Vous pouvez comparer ces deux chaînes.

Nous avons une troisième catégorie. La catégorie similaire, donc c'est dans cette catégorie. Cela prête à confusion par une simple majorité d'utilisateurs attentifs lorsqu'ils ne regardent pas l'une à côté de l'autre ces deux chaînes. Donc, dans la catégorie deux, vous vous rappelez, c'était une vaste majorité. Et il y a des cas pour le latin ici -- l'écriture latine.

Et vous avez les accents. Vous avez les codes ASCII qui sont indiqués également sur les accents graves et aigus sur le U. Il y a des raisons historiques pour utiliser des caractères qui ne sont pas ASCII pour remplacer certaines lettres. C'est principalement dans l'écriture latine que l'on fait cela. Par exemple, mettre un accent différent si, sur une machine à écrire ou sur un clavier, cet accent n'était pas disponible. Donc les exemples sont en écriture latine et en bengali. Donc ça, c'est la catégorie 3. Donc vous les voyez à l'écran et vous voyez un petit peu ces similarités et différences.

Donc nous avons la quatrième catégorie, où il y a assez de similarités distantes. Donc c'est assez similaire, mais on ne peut pas totalement les confondre. La confusion, ce ne serait que pour une minorité d'utilisateurs attentifs ne les observant pas l'une à côté de l'autre. Il n'y a pas véritablement de confusion à ce niveau. Et les experts qui observent ces cas ne pensent pas qu'il y a assez de similarités pour que ces cas soient placés dans les catégories 1,

2 et 3. Donc vous avez deux exemples à l'écran en hébreu et en langues diophantiennes.

Nous allons passer à notre dernière série, la catégorie 5, donc distincte. Ça, c'est la vaste majorité de toutes les paires. Il n'y a pas besoin de les observer, de les analyser. Il y a une forte distinction. Cela ne prête pas à confusion. Et il n'y a pas d'exemples qui sont donnés à l'écran parce que la plupart des paires tombent dans cette catégorie.

Donc, en ce qui concerne le statut de tout cela, eh bien, nous avons, au niveau de ces données et de collecte de données, pratiquement terminé notre travail. Nous avons une préparation pour les commentaires publics. Et nous avons quelques scripts qui sont dans un programme de révision interne et nous avançons dans cette révision interne. Il y a deux scripts qui sont un petit peu en retard. Il y en a un qui est en phase de préparation et c'est pour le thaana. C'est très récemment que le thaana a été rajouté à la zone racine.

Et nous avons également des experts travaillant aux données initiales -- pour des données initiales. Et ça, c'est un cas d'étude pour le Myanmar, où il y a eu quelques problèmes. Il y a eu quelques retards au niveau du Myanmar.

Diapo suivante, s'il vous plaît. Très bien. Donc voilà, ceci conclut ma présentation. Si vous avez des questions, n'hésitez pas. Merci beaucoup.

SARMAD HUSSAIN

Nous allons prendre quelques questions. Nous devons avancer un petit peu dans la présentation. On prend donc -- on prendra les questions à la suite des présentations. Nous allons maintenant entendre de vous.

[TAPANI TARVEINEN]

Une clarification. Vous avez parlé des remplacements ? Est-ce que ça, c'est des cas pour plusieurs caractères. Par exemple, l'umlaut en allemand ?

MICHAEL BAULAND

Eh bien, non. Ce n'est pas inclus ici parce que, ça, c'est pour la similarité. Et on fait très souvent ce remplacement. Et visuellement, ce n'est pas une similarité.

ABDALMONEM GALILA

Merci, Dr Sarmad. Donc vous avez parlé de cinq catégories. C'est une bonne chose, mais vous voyez si un des termes fait partie de l'autre (.god et .godi). Ça, c'est un exemple : Google — G-O-O-G-L-E et G-O-O-O-L-E. Il y a d'autres trucs similaires pour le terme Google. Comment vous allez gérer cela ? Ou est-ce que ça ne doit pas être géré par vous ? C'est tout simplement une différence totale. Comment vous voyez cela ?

Donc, la partie de collecte de données dont j'ai parlé, ce n'est pas au niveau du mot. Ce n'est pas au niveau des étiquettes. Là, c'est

des points. C’est des conjonctions. Par exemple, avec d’autres points code. Ce n’est pas au niveau du mot que l’on travaille. Si vous voulez, c’est des paires qui sont analysées au niveau des similarités. Donc on utilise des données qui s’appliquent aux étiquettes et aux chaînes. Ce n’est pas au niveau du mot.

MICHAEL BAULAND

Donc, la partie de collecte de données dont j’ai parlé, ce n’est pas au niveau du mot. Ce n’est pas au niveau des étiquettes. Là, c’est des points. C’est des conjonctions. Par exemple, avec d’autres points code. Ce n’est pas au niveau du mot que l’on travaille. Si vous voulez, c’est des paires qui sont analysées au niveau des similarités. Donc on utilise des données qui s’appliquent aux étiquettes et aux chaînes. Ce n’est pas au niveau du mot.

SARMAD HUSSAIN

Donc, vous avez donc diverses étapes. Et, c’est ça, le débat que nous avons là-dessus.

ABDALMONEM GALILA

Mais pour valider des numéros mobiles, par exemple, lorsqu’il y a plus de 50 %, on va avoir un problème à un certain niveau. On doit vérifier, et on trouve qu’il y a un numéro qui a été échangé. C’est de cela que je parle. Merci.

SARMAD HUSSAIN On va passer à la dernière partie de cette présentation avec Julian Bernard de Cofomo Quebec, qui veut nous parler d'une mise à jour sur la sécurité.

JULIAN BERNARD Donc j'aimerais vous parler de ce processus d'évaluation des similarités des chaînes. Donc nous avons l'outil SSE qui est donc configuré avec les données SSE. Michael en a parlé tout à l'heure. Et nous configurons cela avec les différents noms des TLD ccTLD non bloqués, et les noms, également, que nous voulons comparer avec d'autres chaînes.

Donc cet outil, nous avons reçu -- dans le cadre de TAMS, du système de gestion de candidature des gTLD. Nous avons reçu donc ces données et nous avons des rapports qui sont générés avec toutes les chaînes qui sont demandées. Et ça, ça va nous permettre, avec cet outil SSE, de voir les rapports. Et si un panéliste pense qu'il n'y a pas assez de données pour prendre une décision, eh bien, il sera en mesure de retravailler sur la chaîne et d'assembler un rapport final, de continuer à travailler à cela.

Donc l'outil -- vous avez donc la possibilité de générer des rapports personnalisés. Et le panéliste sera donc en mesure d'éditer le résultat avant de rajouter cela au rapport final. Et il va donc pouvoir créer donc ce rapport pour évaluer donc les sous-ensembles de chaînes pour lesquels il y a eu des demandes. Et ensuite, nous aurons donc — le responsable de panel — des informations qui seront fournies. Et dans le cadre du rapport final, nous serons en

mesure de télécharger ce rapport final dans le système de gestion de candidature des gTLD, TAMS, qui va déterminer, par exemple, si une chaîne pour laquelle on a fait une demande est similaire à une autre. S'il y a donc un ensemble conflictuel éventuel.

Et ça va être la même chose avec les similarités avec les noms bloqués. Et on va voir donc s'il y a un rejet qui sera nécessaire à ce niveau.

Donc je regarde que je n'ai rien oublié sur ce transparent. Je voulais donc confirmer que cet outil est donc pour les panélistes, pour les aider à prendre des décisions. C'est un guide si vous voulez, cet outil. Et ça permet d'améliorer la qualité des rapports qui vont être effectués.

Nous allons passer à la diapo suivante. Alors l'outil peut être accessible sur icann.org. Si vous avez un compte ICANN, donc il sera accessible pour les membres du panel, et il fournira l'analyse automatisée de similarité de chaînes. L'objectif étant, bien sûr, d'optimiser la révision manuelle par le panel.

Les données de similarité des chaînes décrites par Michael seront incorporées, bien sûr, dans cet outil. Et les chaînes faisant l'objet d'une candidature seront comparées à la liste des chaînes que Sarmad a décrite plus tôt. Donc, toutes les chaînes faisant l'objet d'une candidature, les gTLD ou les ccTLD existants, les IDN ccTLD requis, les gTLD faisant l'objet d'une candidature dans le cadre de

la série précédente, les caractères ASCII à deux caractères et les non bloqués.

Donc, cet outil passera par trois étapes. Tout d'abord, il détectera les étiquettes qui pourraient prêter à confusion. Ensuite, il calculera et déterminera un score de similarité ou de distance visuelle. Et lorsque les notes de similarité dépassent un certain seuil, elles seront incluses dans le rapport par l'outil, rapport qui sera soumis donc aux membres du panel afin qu'ils puissent le réviser de manière manuelle.

Ce seuil est donc une valeur configurable. Ce seuil est conflictuel, donc il permettra de déterminer la sensibilité de la comparaison. Donc s'il y a un score de similarité qui est inférieur au seuil conflictuel, alors l'outil déterminera -- inclura plutôt dans le rapport les variantes qui disposeront de la note de similarité la plus élevée.

Alors ces rapports émis par le système seront basés sur les chaînes faisant l'objet d'une demande. Il sera généré ensuite en utilisant des valeurs de seuil prédéterminées. Et les rapports personnalisés seront quant à eux établis par les membres du panel. Et les membres du panel, donc, pourront déterminer une valeur de seuil conflictuelle personnalisée. Ces rapports seront disponibles en format CSV ou HTML. CSV, valeurs séparées par des virgules, et le format HTML présentera le rapport de façon à ce qu'il soit présenté aux membres panélistes [qui ne prêtent, qui ne posent aucun problème] comparé au modèle CSV.

Quel est le mécanisme de base de cet outil ? Eh bien, sur base des données, cet outil va calculer si une chaîne faisant l'objet d'une candidature est similaire à une autre chaîne faisant l'objet d'une demande à des TLD existants, des IDN ccTLD requis, des ASCII à deux caractères ou des noms bloqués ou leurs variantes. Et donc ces chaînes similaires sont regroupées dans des ensembles conflictuels, dans des potentiels ensembles conflictuels. Et pour chaque ensemble conflictuel, on va calculer une note, un score pour chaque paire de chaînes -- pour chaque paire de variantes. Ensuite, le rapport qui sera soumis aux membres du panel afin qu'ils puissent effectuer leur révision finale contiendra des chaînes potentiellement similaires, avec un score en deçà d'un certain seuil.

Et donc, lorsqu'on va rassembler les variantes, les chaînes, on va obtenir un score. On va comparer donc les variantes et les chaînes. Et pour une paire de chaînes, on va en fait choisir la chaîne qui aura obtenu le score le plus bas. Et l'indicateur permettant de calculer ce score de similarité sera déterminé après l'analyse des données.

Diapo suivante, s'il vous plaît. Le rapport final contiendra cette liste d'éléments. Donc certains ensembles conflictuels, les identifiants uniques, les étiquettes et les identifiants d'application de l'étiquette qui fait l'objet d'un conflit. L'étiquette 2 est la source de cette deuxième étiquette. Ensuite, les variantes des étiquettes qui seront considérées comme étant les plus similaires dans le cadre

de cette comparaison. Puis ensuite, on aura le score. Le niveau de similarité n'atteindra que 3 ou sera inférieur à 3.

Diapo suivante, s'il vous plaît. Donc voilà. Ici, vous avez à l'écran un exemple de ce à quoi ressemblera ce rapport. Pour cet exemple théorique, on a un nom bloqué qui est pris, donc qui est « example ». Et puis on a des chaînes similaires incluses dans le rapport. Par exemple lorsque le « m » est remplacé par les lettres « rn ». Un « I » qui, lorsqu'il est en majuscule, ressemble à un « l », et les accents aussi sur les « I » qui sont des variantes prises en compte.

C'en est tout pour moi. Y a-t-il des questions ?

SARMAD HUSSAIN Non, je ne vois pas de mains levées dans la salle ni sur Zoom. Donc passons maintenant à l'intervention de David afin de parler de notre calendrier. Et puis nous vous rendrons la parole si vous avez encore des questions, cher public.

DAVID WAGHALTER Merci. Donc je suis David Waghalter, gestionnaire produit au sein du groupe IT de l'ICANN. Et nous travaillons à cet outil qu'on vient de vous décrire.

Voilà notre calendrier à l'écran. Actuellement, la phase de développement est en cours. Nous avons déjà respecté la plupart des critères à respecter pour cette étape. Nous sommes en train de tout peaufiner, de tout finaliser, et nous allons poursuivre le

développement de cet outil lors du mois d'avril. Ensuite, nous aurons une UAT interne pour parler du MVP, donc du produit qui représente la plus grande valeur. Et puis nous tenterons de déterminer certaines exigences initiales une fois qu'on aura évalué tout ceci en interne. Ensuite, en juin, nous développerons des critères et des exigences basés sur le résultat de notre travail, et ce travail se poursuivra durant le cours de l'été. L'objectif étant d'avoir une session questions-réponses en septembre, et puis de publier officiellement cet outil en septembre.

SARMAD HUSSAIN Bien, merci. Y a-t-il des questions ? Merci David. Alors cela nous amène à la fin de notre présentation. Je vais maintenant vous donner la parole afin que vous puissiez poser vos questions si vous en avez ou, bien sûr, nous faire part de vos commentaires.

VINNY ADJIBI Bonjour et merci pour cette présentation. Je voulais vous demander si les données utilisées pour cette évaluation seront rendues accessibles, disponibles.

SARMAD HUSSAIN S'il vous plaît, présentez-vous lorsque vous prenez la parole.

VINNY ADJIBI Oui, je suis participant Nextgen.

SARMAD HUSSAIN

Oui. Alors nous avons l'intention de publier les données que nous recevrons de la part des experts pour les soumettre à un commentaire public. Et puis ensuite, une fois qu'on aura pris en compte les commentaires de la communauté, nous finaliserons les données qui seront utilisées par cet outil. La période de commentaires publics devrait commencer dans les deux ou trois semaines à venir.

Alors je vais peut-être vous donner encore une minute pour réfléchir à vos potentielles questions. Pendant que vous réfléchissez, je voudrais encore une fois répéter que cet outil est encore en cours de développement. Cette phase est en cours et, à l'heure actuelle, nous ne pouvons pas vous garantir que cet outil sera accessible à la communauté, car il est développé pour qu'il soit utilisé par le panel d'experts. Une fois que ce sera le cas, nous parlerons de le rendre accessible ou non.

STEVE DELBIANCO

Hier, nous avons eu une réunion ou une présentation de la part des vendeurs qui aident les bureaux d'enregistrement et qui bloquent de nouveaux noms de domaine qui sont utilisés à des fins malveillantes par exemple pour des raisons de fraude. Et ces vendeurs -- ces fournisseurs ont indiqué une efficacité accrue de la part des outils d'IA qu'ils utilisent. Mais il semblerait qu'ils dépendent notamment de l'historique du bureau d'enregistrement qui indiquerait peut-être qu'il y a là l'intention de duper les

utilisateurs. Avez-vous envisagé d'inclure des outils IA dans votre outil ? Et peut-être bien sûr que les experts membres du panel voudront décider d'utiliser ou non. Voilà.

Si vous l'avez déjà dit, pardonnez-moi de reposer cette question, mais voulez-vous utiliser l'IA dans le cadre de votre outil ?

SARMAD HUSSAIN Non, l'IA ne sera pas utilisée. Cet outil est basé notamment sur le score de similarité qui sera fourni par les experts. Mais je sais au moins que, pour un script, certaines des données collectées et utilisées ont été générées par un outil d'apprentissage machine et elles ont été utilisées pour déterminer des points de code similaires. Cela a été fait pour le script han, donc le script chinois, étant donné qu'il contient énormément de caractères différents. Donc là, oui, l'IA est utilisée pour déterminer le caractère potentiellement similaire de certains caractères. Et bien sûr, nous partagerons ces détails lorsque les données seront publiées.

STEVE DELBIANCO Oui, par IA je voulais vraiment dire IA et non pas apprentissage machine. Donc par exemple, une IA générative basée sur de grands modèles de langue, de langage. Si en septembre, on publie un outil qui ne reconnaît pas la valeur ajoutée que pourrait apporter l'IA, alors que les autres utilisateurs vont utiliser l'IA pour évaluer les décisions prises par cet outil, que va-t-il se passer ? On est en 2025.

On est à l'ICANN. Je ne sais pas. Peut-être, utiliser au moins un LLM gratuit dans le cadre de votre processus.

SARMAD HUSSAIN Merci pour ce commentaire. Nous en tiendrons compte.

ABDALMONEM GALILA Alors certaines des données qui sont déjà disponibles, peut-être qu'elles sont basées sur des jeux de données. Mais je pense que tout dépend aussi des points de code quand il s'agit de similarité de chaîne. Donc peut-être que c'est différent si l'IA dépendait des points de codes. Et j'en doute. Peut-être que ce serait utile. Sinon, je ne pense pas qu'il serait judicieux d'utiliser l'IA à cette fin.

SARMAD HUSSAIN Merci beaucoup Abdalmonem. Quoi qu'il en soit, nous avons bien pris note de votre commentaire. Je vois que nous avons une autre main levée. Vanda sur Zoom, je vous en prie.

VANDA SCARTEZINI Oui, bonjour. Je voulais juste vous informer du fait que, au sein de notre registre, si quelqu'un veut enregistrer un nom de domaine IDN, donc internationalisé, qui contient des caractères similaires, par exemple la lettre O dans le code ASCII ou la lettre O dans la langue arménienne, eh bien, ces caractères sont similaires visuellement.

Et on ne peut pas enregistrer un IDN en utilisant une autre police. Si le premier O du mot provient du code ASCII, alors on peut l'enregistrer, ce nom de domaine IDN. Voilà, je voulais juste vous informer de cela, car cela pourrait aider à détecter les sites d'hameçonnage sur base de cette similarité.

SARMAD HUSSAIN Merci Vanda. Alors nous avons respecté le délai qui nous était imparti. S'il n'y a plus d'autre question ou commentaire -- je ne vois pas de mains levées. Dans ce cas, je voudrais vous remercier d'avoir participé à cette réunion. Nous aimerions bien sûr aussi remercier les membres du panel pour leur intervention et nous souhaitons maintenant lever cette séance. Et l'enregistrement a pris fin.

[FIN DE LA TRANSCRIPTION]