ICANN78 | Reunión General Anual – Planificación de una revisión de similitudes entre cadenas de caracteres en la próxima ronda de nuevos gTLD
Jueves, 26 de octubre de 2023 – 9:00 a 10:00 HAM

PITINAN KOOARMORNPATANA: Hola. Bienvenidos a la sesión de planificación de la revisión de similitudes de cadenas de caracteres en la próxima ronda de gTLD. Soy Pitinan y coordinaré la participación remota. Tengan en cuenta que esta sesión está siendo grabada y que se rige por los estándares de comportamiento esperados de la ICANN.

Esta sesión incluye interpretación y transcripción en tiempo real que pueden habilitar al hacer clic en los iconos correspondientes de la barra de Zoom. Si desean tomar la palabra, por favor, levanten la mano y el moderador les dará la palabra. También pueden colocar su pregunta en el chat. Todos los participantes deberán mencionar su nombre y hablar a una velocidad razonable. Ahora le voy a dar la palabra a Sarmad. Adelante, por favor.

SARMAD HUSSAIN: Muchas gracias, Pitinan. Buenos días a todos. Buenos días, buenas tardes y buenas noches a todas las personas que están conectadas remotamente. Esta sesión es la primera en una serie de sesiones en las cuales nos vamos a centrar en la revisión y la planificación de la revisión de las similitudes de cadenas de caracteres en la próxima

ronda de gTLD. A mi lado tengo a Asmus Freytag, quien nos va a contar sobre las pautas que hemos estado desarrollando junto con Michelle DeSmyter. Ambos están trabajando en este proyecto con nosotros.

Antes de pasar a las generalidades de la presentación en sí pensé que quizá sería de utilidad darles un poco de contexto sobre algunas cuestiones o caracteres similares y diferentes porque son cuestiones que abordaremos durante nuestra conversación del día de hoy. Cuando hablamos de cadenas de caracteres o etiquetas de TLD que se solicitan o que se utilizan las mismas se componen de caracteres o puntos en Unicode. Desde un punto de vista visual pueden verse como similares o distintas entre sí. En algunos casos se combinan para poder ser consideradas visuales similarmente o distintas.

Se pueden agregar dos capas. El carácter en sí, que es lo que compone la cadena de caracteres a nivel de la cadena de caracteres. Algunas de estas cadenas, cuando se combinan los caracteres cambian la forma. Hay que tomar en cuenta los caracteres individuales. Analizarlos por sí solos no es suficiente. Necesitamos brindar cierta información. Para la revisión de la similitud de las cadenas es necesario analizar toda la cadena.

Esto es simplemente para darles algo de contexto. Aquí hay algunos ejemplos de caracteres en Unicode, en puntos de código de algunos códigos de escritura diferentes para que puedan observar las categorías. La primera regla tiene que ver con los que son iguales a nivel visual. Básicamente se representan esos caracteres de quienes utilizan esos códigos de escritura y que pueden no ser distinguibles

entre sí. En algunos casos hay una pequeña diferencia pero no se puede discernir con facilidad. En otros casos son exactamente iguales en cuanto al formato. Se utiliza el mismo glifo.

Si observan el ejemplo en cirílico y latín, el carácter C y el cirílico que es también C pero que se representa como 0441, son similares. Lo mismo ocurre con el código de escritura árabe. Verán entonces que el carácter que está representado en 0649 y 06CC son idénticos. Esto nos brinda algo de contexto.

Estos son ejemplos de lo que nosotros denominamos cadenas que son iguales desde el punto de vista visual. Aquí estamos haciendo una interpretación visual y ningún otro tipo de interpretación. Caracteres que son visualmente similares son identificados por la comunidad de los códigos de escritura y los que están trabajando en las reglas de generación de etiquetas. Esto básicamente nos da un ejemplo de dos caracteres, quizá en latín y griego, que pueden ser considerados similares. Tenemos un ejemplo de Devanagari y Bangla, que se pueden considerar similares y otros caracteres también que se pueden distinguir entre sí. Por ejemplo, la F en inglés u otros caracteres en otros códigos de escritura. La primera fila representa caracteres que parecen ser iguales y han sido identificados por las comunidades de códigos de escritura y ya han sido establecidos como variantes en la generación de etiquetas. Ya han sido codificados y ya han sido abordados.

La revisión de las cadenas de caracteres y su similitud tiene que ver con analizar caracteres que son visualmente similares y también los

códigos de escritura que pueden ser similares y dependerá de los paneles de revisión que se configurarán para la próxima ronda si estas cadenas quedan dentro de esta categoría, es decir, si son similares visualmente o si crean confusión a nivel visual. En ese caso no van a poder proceder porque van a crear confusión. No van a poder delegarse. Si son similares pero se pueden distinguir, entonces sí pueden avanzar en el proceso.

Hay una distinción que el panel de revisión tiene que hacer en este sentido. De esto se trata el proceso de revisión de similitudes de códigos de escritura. Estamos trabajando en este sentido para establecer o determinar los detalles de lo que sería un buen mecanismo para avanzar con este proceso y que este proceso sea más robusto.

Pasemos ahora a la presentación. Lo que haremos en esta sesión es lo siguiente. Vamos a presentar cierto contexto y hablaremos de las recomendaciones que surgieron de las políticas. El primer conjunto de recomendaciones viene del proceso de desarrollo de políticas del SubPro y además existen más detalles que se han agregado en relación a la similitud de cadenas de caracteres.

En las recomendaciones de la fase uno del proceso de desarrollo de políticas expedito en relación a los nombres de dominio internacionalizados o IDN EPDP se mostró lo siguiente. Por favor, tengan en cuenta que este informe está en su fase uno. Son las recomendaciones de la fase uno que están para comentario público. Lo que compartimos aquí son las recomendaciones que están

sometidas a comentario público. No son las recomendaciones finales del EPDP. Están en el comentario público. Se están revisando algunas de esas recomendaciones y las recomendaciones finales aún no han sido publicadas.

Los detalles que vamos a compartir sobre el EPDP de IDN probablemente pueda cambiar pero a los fines de esta presentación estamos utilizando el PDP que se utilizó para el comentario público. Teniendo en cuenta esto vamos a revisar esas recomendaciones y vamos a brindar un alcance. También determinaremos la metodología y los resultados de la revisión de la similitud de cadenas de caracteres. Como ustedes pueden ver, estas recomendaciones requieren el desarrollo de pautas de revisión de similitud de cadenas de caracteres y ese es el trabajo de este panel.

Asmus nos está ayudando también a desarrollar estas pautas que finalmente se usarán en este proceso. Estas pautas aún no están finalizadas. Hemos comenzado a trabajar en las mismas y lo que haremos es avanzar en el proceso de confección y también buscaremos sus aportes para ver si estamos avanzando en la dirección correcta o si debemos hacer algunas modificaciones al proceso al momento de desarrollar estas pautas. Esto es lo que abordaremos en esta sesión.

Comencemos entonces teniendo en cuenta las recomendaciones de políticas. La sección 24 del SubPro habla de la revisión de la similitud de cadenas de caracteres. El grupo de trabajo afirma la recomendación 2 de la política de 2007 que señala que las cadenas de caracteres no

deben ser confusas o similarmente confusas en un TLD existente o en un nombre reservado. Luego tenemos la afirmación 24.2, que hace referencia a la ronda de 2012 de los nuevos gTLD y que brinda más detalles sobre estas comparaciones y el alcance y, lo más importante, nos da una especie de definición de lo que significa similarmente confusa para una cadena de caracteres. Similar significa que las cadenas de caracteres son tan similares que pueden crear una probabilidad de confusión del usuario si más de una cadena de caracteres se delega en la zona raíz.

Básicamente el panel de similitud de cadenas de caracteres tuvo la tarea de determinar cuál es la diferencia entre una similitud simple versus una similitud que puede causar confusión al usuario final. Esos son los detalles que hay que tener en cuenta.

También hay otras recomendaciones adicionales que hablan de una herramienta automática que pueden no ser suficientes para avanzar con el proceso y también se habla de algunos detalles del proceso de objeción. En las recomendaciones de la fase uno del EPDP hay muchos más detalles en relación a la revisión de la similitud de cadenas de caracteres. La recomendación 4.1 claramente delinea el alcance donde dice que las cadenas solicitadas deben ser comparadas con gTLD existentes, ccTLD existentes, otras cadenas solicitadas como ccTLD con IDN, todas las cadenas de gTLD que fueron solicitadas, los nombres reservados y también las cadenas con dos caracteres en ASCII.

Esto sugiere que estas cadenas de caracteres van a ser comparadas con las variantes asignadas y que estas variantes también van a ser comparadas con otros bloques de variantes. Básicamente esto brinda mucho detalle con respecto a las comparaciones. Esto es en lo que respecta al alcance de las comparaciones y lo que se refiere a las recomendaciones.

También este alcance de comparación puede ser muy extenso porque se comparan las cadenas con las variantes, con otras cadenas y con otras variables y existen cadenas de caracteres que pueden tener cientos de variantes. Por lo tanto, el resultado es que se compararían cientos de cadenas de caracteres con cientos de variantes.

En algunos casos, estas comparaciones pueden ser muy extensas y probablemente no determinen si existen temas de confusión pero sí nos permite tener cierta eficiencia en el proceso, pero solamente hasta un punto. Si uno por ejemplo compara chino y árabe, quizá la probabilidad de similitud es muy baja entre estos dos códigos de escritura pero si comparamos el cirílico y el latín, la probabilidad de similitud va a ser muy alta.

Se deben desarrollar pautas muy detalladas con respecto a lo que se debe comparar y estas pautas y criterios deben también identificar qué es lo que se podría saltear cuando hay bajas probabilidades de similitud en diferentes pares de códigos de escrituras. Estas pautas en particular dan lugar a estas excepciones pero son pautas que se deben desarrollar antes de avanzar.

Finalmente, esta recomendación señala que si hay cuestiones con la similitud en un grupo o conjunto de variantes, todo el conjunto de variantes queda impactado, no solamente la cadena de caracteres. Si una cadena de caracteres es similar a otra y esto se da en la próxima ronda de los gTLD, entonces la totalidad de la variante y la cadena de caracteres va a pasar a un grupo contencioso, no solamente la cadena. Si las variantes coinciden, aunque no sean similares exactamente, también pasarán a este conjunto contencioso. Esto nos da un contexto con respecto a la política y por eso se solicita desarrollar una metodología, un alcance y resultados.

Volviendo al alcance, como ustedes podrán ver lo que se sugiere es, en la parte superior, aquí vemos las cadenas de caracteres de gTLD que han sido solicitadas. Esta es la cadena primaria en sí y luego puede tener algunas variantes asignables y también etiquetas de variantes bloqueadas. Esto depende de las reglas de generación de etiquetas.

A la izquierda verán las categorías de comparación que se han identificado en el EPDP. Esto incluye los gTLD existentes o gTLD que han sido solicitados pero que todavía se encuentran en proceso de la ronda anterior. Es decir, son cadenas de caracteres que no están procesadas todavía de la ronda del 2012. CcTLD existentes o TLD con código de país existentes. Hay algunos que están todavía en el proceso de solicitud y que no han sido delegados todavía. Hay otras cadenas de gTLD solicitadas, los nombres reservados y también los nombres de dos caracteres en ASCII, que son gestionados por la ISO.

Básicamente van a tener una restricción de tiempo asignado a la variable en las reglas de generación de etiquetas y todas estas variantes tienen que ser comparadas con las cadenas de caracteres y variantes de etiquetas solicitadas. La palabra yes, sí, indica las requeridas. Donde dice yes con asterisco son las áreas especificadas. Estas son las comparaciones en las que el panel puede decidir omitir, como decía ayer, si comparamos el chino con el árabe, el panel puede decidir que las posibilidades de similitud son muy bajas y, por lo tanto, para una cadena en particular, decidir omitir esas comparaciones. Esas son las que tienen indicado el yes con el asterisco. Las variantes en sí no pueden omitir todas las comparaciones sino que se refieren solamente a algunas de ellas. Donde dice no, se trata del área donde no hay un ámbito de comparación para similitud de cadenas de caracteres y, por lo tanto, no se pueden comparar. Este es un resumen de lo que se ha establecido en la política, la recomendación 4.1 del EPDP para IDN.

La metodología implica que al principio tenemos que compilar una lista de todas estas cadenas de caracteres. Tienen que provenir de todas las diferentes categorías y debemos compilar una lista de las variables asignables y posiblemente las variantes bloqueadas. Las listas de variantes bloqueadas puede ser amplia para una cadena en particular. No en general pero para una cadena en particular sí puede ser más grande, con lo cual deberemos establecer cuál es el mecanismo más viable para hacer este tipo de comparaciones.

Por supuesto que los detalles estarán una parte en los lineamientos y otra parte estarán los determinará el panel en sí. Una vez que se hayan identificado todas las cadenas de caracteres, el panel es quien decidirá cuáles son las comparaciones que omitir o qué bloqueos omitir. Luego se identificarán las distintas aplicaciones con la misma cadena de caracteres contenciosa. Si hay dos aplicaciones diferentes, pueden ir a contención por esto o también ocurrirá cuando haya dos cadenas de caracteres que son iguales. Para el resto se hará una comparación de cadena de caracteres en la medida en que la confusión, la definición de lo que es confusamente similar pueda provenir de las normas, los lineamientos de la similitud.

El panel hará un documento que identifique los sets contenciosos de la similitud y esto provendrá del panel de la ICANN y los resultados se publicarán. Esto es un panorama general del proceso que se seguirá en base a los lineamientos que nos establezcan las recomendaciones. Los resultados, por supuesto, hay varios resultados posibles. Habrá una revisión del panel. Podrá haber una cadena de caracteres que sea similar a un gTLD existente que debe ser documentado y en ese caso quizá no proceda. Hablaremos de eso en la próxima diapositiva. Se puede documentar que una cadena es similar a una cadena de gTLD solicitada en la ronda anterior o incluso similar a un ccTLD o a un IDN con ccTLD o a un nombre reservado o a dos caracteres. Lo que no está incluido en esta lista es que puede ocurrir que sea similar o una variable asignable o una variable bloqueada. Estas son opciones que discutimos en la diapositiva anterior y si cualquiera de estas tiene este resultado puede estar en esta lista o si se encuentra que esta cadena

de caracteres no es similar a una de estas categorías, con lo cual se puede proceder. La organización de la ICANN publicará todo esto para revisión, para que todos puedan ver el proceso.

Aquí vemos un poco más de detalle. No voy a mencionar todo pero estos Resultados tienen consecuencias. Para tomar un ejemplo de la izquierda superior, ahí vemos la primera regla. Si la variante solicitada para el gTLD o cualquiera de las cadenas de caracteres se encuentra que es parecida a un gTLD existente, no puede proceder. Si se encuentra una variable similar a un gTLD existente, no puede proceder hasta que sea una variable asignable y se aplica al mismo operador de registro, con lo cual es una variable. Si es similar pero no una variante de un gTLD existente, en ese caso tampoco puede existir.

Pueden ver en las diapositivas que hay otro conjunto de consecuencias del panel de revisión de similitudes. Claro que esto nos va a guiar en el desarrollo del proceso en sí. Ahora sí, le voy a dar la palabra a Asmus, quien nos va a mencionar la guía en sí, que en este momento está siendo desarrollada.

ASMUS FREYTAG: Gracias, Sarmad. Usted cubrió prácticamente todo lo que yo iba a decir. Vamos a la próxima diapositiva. Vamos a reiterar que la confusión es la medida en la que dos etiquetas pueden ser sustituidas una por otra y la sustitución es un concepto clave aquí. Cuando hablamos de confusión hablamos de algún tipo de umbral que pueda existir que no se puede definir objetivamente pero hay que elegir uno y

ahí decimos que la confusión ha logrado un nivel en el que importa mucho para la zona raíz porque la confusión o la confundibilidad existe en un amplio espectro. La similitud de cadenas de caracteres es un conjunto que tiene una apariencia similar y por lo tanto podemos tener el concepto de una confusión similar donde el nivel de similitud que necesitamos excede ese umbral que no está definido donde la confusión ya no es tolerable.

Resulta muy difícil, y esto es central, establecer una norma, un estándar de confusión porque lo que se ve confuso es distinto para cada usuario y también para el contexto donde se presenta esta etiqueta. La gente que está familiarizada con un idioma o una cadena de caracteres en un código de escritura específico ha pasado años y años y años leyendo cadenas de caracteres en un código de escritura y esto es algo que puede no ser compartido con personas que están familiarizadas con esa misma lengua.

Para aislar unas directrices hay que poder definir cuál es el usuario meta. Lo primero que vamos a sugerir es que el usuario al que estamos apuntando es lo que se denomina razonablemente atento. Si no estamos mirando en qué estamos haciendo clic vamos a confundir todo. Si se trata de un topógrafo, usted va a hacer distinciones que no hace otra persona. Se trata de un término medio de sentido común.

El otro requisito para ser razonable es que cuando se trata de direcciones de Internet y cadenas de caracteres, apuntamos a personas que están familiarizadas con el código de escritura o el idioma en el que la etiqueta está presentada. Esto significa que no

podemos descartar una situación donde un usuario está buscando una etiqueta en el idioma nativo o en el código de escritura pero sin embargo se le presenta un script que es ajeno al usuario y que tiene una cadena de caracteres que de todos modos es similar.

Lo que sí excluiríamos de esta definición meta de un contexto meta es una situación en la que un usuario tiene que decidir entre dos etiquetas en un código de escritura que un usuario no puede leer. Si tratamos de establecer cómo es una similitud que causa confusión entre dos cadenas de caracteres por parte de un usuario, esta es una pregunta que no podemos responder porque se trata de que una etiqueta en chino se ve como una etiqueta en chino. Entonces estamos suponiendo que una parte significativa de los usuarios de facto están familiarizados con el subconjunto ASCII y también podemos asumir que muchos de ellos quieren encontrar la etiqueta ASCII correcta para acceder al recurso. La definición de un usuario objetivo tiene que cubrir el contexto donde un nativo en un código de escritura diferente está buscando una etiqueta ASCII.

Aquí debería haber otra viñeta. Estamos excluyendo el contexto donde el usuario está tratando de distinguir entre dos conjuntos de etiquetas que visualmente están presentes al mismo tiempo. Esta es una situación que se usa muy comúnmente para los tests artificiales o los tests de confusión pero aquí se trata de que el usuario tiene una idea del nombre y entonces busca la corporación física de ese algo y debe decidir si ese algo coincide con aquello que está buscando. Tenemos

un caso donde hay dos etiquetas una al lado de la otra y hay que elegir una.

El concepto clave aquí es que cuando tratamos de hacer una revisión de similitud de cadenas de caracteres lo que importa al final no es que los puntos de código son iguales sino cuál es el resultado de la cadena de caracteres. Es decir, que sea similar o confundible. Una de las cuestiones cuando tenemos puntos de código confundibles es si una etiqueta contiene una subcadena de caracteres que está compuesta de puntos de código que no tienen ningún equivalente similar en ningún otro lado. En ese caso, debería ser suficiente poder hacer que la totalidad de la etiqueta sea única, porque no se puede construir algo que se confunda con esa etiqueta incluso si hay otra parte de la etiqueta que tiene un equivalente de confusión.

Otro aspecto interesante es que cuando la gente mira las etiquetas las trata como un reconocimiento de palabra o una tarea de lectura normal y sabemos que el reconocimiento de palabra es holístico. La gente no lee las palabras letra por letra sino que lee la palabra entera. Esto significa que si una suficiente cantidad del mundo tiene un contexto donde subconscientemente se dispara el reconocimiento por parte del usuario a pesar de que hay ciertos detalles que están mal, para la lectura holística se pueden superponer algunos detalles y presentaciones a pesar de que no se trata de un caso típico o sí hay un caso típico.

Hay muchos lectores en inglés que simplemente no reconocen que hay un diacrítico arriba o abajo de una letra para crear una etiqueta falsa.

En la lectura de ellos, esa marca no existe. No las identifican. Les parece que es una suciedad en la pantalla. En ese caso pueden hacer clic en la etiqueta equivocada porque hay un acento donde no debería estar. Creo que he cubierto todo. Vamos a la siguiente diapositiva.

Vale la pena distinguir la similitud del concepto de variante. Con las variantes definimos una etiqueta que de alguna manera es tratado de manera igualitaria que la etiqueta original. Es decir, si tiene la misma apariencia o es tan cercano a esta apariencia que el usuario piensa que es lo mismo. La variante no está limitada a la similitud visual, que es un concepto muy importante. Hay un caso muy común donde tenemos la forma de chino simplificada y tradicional, y en este caso estos caracteres se consideran el mismo carácter. Los usuarios podrían no tratarlos conscientemente como diferentes a pesar de que no son similares visualmente. En el etiópico tenemos un caso en el que muchas letras producen el mismo sonido y son sustituidas por la gente que escribe el idioma. No son similares pero son etiquetas que son tratadas del mismo modo con unas letras escritas de modo diferente.

La relación de variante es una relación equivalente y todas las relaciones equivalentes son transitivas. Esto quiere decir que cualquier variante de una etiqueta también es una variante de todas las otras variantes. Sin embargo, la similitud a nivel general significa que si la diferencia es pequeña y por esa diferencia que es detectable a pesar de que puede ser ignorada por los lectores, es que dos etiquetas pueden ser similares a una tercera etiqueta pero no son lo suficientemente parecidas entre sí para ser similares. La similitud y la confusión no son

transitivas porque se puede eliminar alguna de las etiquetas eficientes que tenemos.

Hay algunas variantes que están basadas en la similitud visual. En ese caso tienen unos pares idénticos o casi idénticos. Una apariencia idéntica o casi idéntica. Una de las cosas que las variantes no consideran es la similitud en mayúsculas. Las etiquetas ASCII pueden estar presentes en mayúsculas o en minúsculas e incluso los IDN son aceptados en mayúscula por la mayor cantidad de los usuarios. Por lo tanto, se pueden tener etiquetas que están presentes en mayúscula y el usuario tiene que definir si esa es la etiqueta que quiere. En la zona raíz no las hemos incluido en la variante por la transitividad. Tengo un ejemplo para mostrarles cómo esto les puede meter en problemas. La similitud de cadenas de caracteres no tiene la restricción de ser transitiva. La confusión en base a la equivalencia en mayúsculas subyacente debe ser el alcance. Siguiente diapositiva, por favor.

Aquí vemos un ejemplo de lo que les mencionaba. Comenzamos con la letra H. En la parte superior izquierda hay una mayúscula H. Todos estamos familiarizados en el conjunto de ASCII con esta letra. Muchos de los usuarios no serían conscientes de si esto es mayúscula o minúscula. El caso de la H mayúscula combina con la mayúscula en griego. Esto normalmente sería una variante excepto que ignoremos las variantes de mayúsculas. Si se tiene una etiqueta que tiene por ejemplo una mayúscula, quizá el usuario no pueda diferenciar si es latín o griego.

En el caso de las minúsculas es distinto pero en algunos casos la similitud con la minúscula es poca y puede confundir. Si las cosas se presentan en el contexto adecuado, se puede entonces prever la ausencia o la presencia de estas diferencias. Las mayúsculas en ASCII a veces tienen la misma apariencia que las mayúsculas en griego. No es una variante porque lo excluimos pero si la etiqueta está presente en mayúscula, no es posible diferenciar si se trata de un carácter latino o griego.

La N griega tiene una minúscula que se representa con 03BD que es muy similar a la v minúscula. En algunos contextos y con algunas etiquetas se tiene un cierto grado de confusión. La minúscula de la v tiene un equivalente en la mayúscula. Si queremos completar el círculo veremos que la mayúscula V y la minúscula h son visualmente muy diferentes y no hay posibilidad de que el usuario las confunda. Por eso estas cuestiones las tenemos que abordar a nivel de las variantes. Tendríamos todas las n, las h y las v en latín y las variantes, lo cual no sería demasiado práctico.

En cuanto a la similitud de cadenas de caracteres, tenemos que incluir estas cuestiones o estos efectos porque caso contrario se podrían crear etiquetas muy confusas para el usuario. Supongan que tenemos una etiqueta es HN. Si está en mayúscula no se puede diferenciar de si se trata de latín o griego. Pasemos por favor a la próxima diapositiva.

Esto repite un poco las métricas y algunas de las cuestiones que ha presentado Sarmad. Dado que todas las variantes son consideradas un código dentro de la escritura se tornan confusas con sus variantes o

con las etiquetas delegadas. Desde un punto de vista técnico esto hay que tenerlo en cuenta porque incluso las variantes que están bloqueadas pueden ser tratadas como algo similar por un mismo usuario. En este caso son dos posibles ejemplos que se excluyen de la comparación. Pasemos a la próxima diapositiva, por favor.

Existen ciertas interacciones que resultan interesantes con las distintas cadenas de caracteres y las vamos a explorar en la próxima diapositiva. Aquí vemos un ejemplo común de cómo interactúan las diferentes cadenas de caracteres. A la izquierda tenemos un carácter latino que muchos de nosotros no necesitamos pero que sí se utilizan en ciertos idiomas. A la derecha tenemos un carácter armenio. Podremos discutir si es similar o distinto. Yo los he mostrado y tienen una muy pequeña superposición. Hay gente que los considera distintos porque uno tiene una curva en una posición distinta pero si agregamos la letra similar griega vemos que tiene una curva de una forma. Existe una gran superposición en relación a la similitud. Dependiendo de si hablamos de que las dos superiores son lo suficientemente confusas para que estén presentes, seguramente no pueden estar presentes al mismo tiempo.

El concepto de similitud es que depende del observador y esto es sumamente importante cuando nos referimos a cadenas de caracteres que son muy diferentes en su naturaleza porque los lectores lo que hacen es buscar características particulares para ver en qué se diferencian. Si bien pueden ignorar ciertas características, hay otras que pueden considerar de importancia.

¿Qué tipografía o qué característica se toma en cuenta de cadena a cadena? Para algunas cadenas de caracteres existen datos disponibles donde ya se sabe qué puede causar confusión. En otros casos no y para algunas cadenas de caracteres determinar lo que puede ser confuso es muy complicado porque hay cadenas con, por ejemplo, caracteres adyacentes que forman ligaduras que son muy similares a las del latín. Creo que todavía se tiene que llevar adelante mucha investigación para poder entender y determinar qué sucede con esas formas que se funden porque a veces analizar la cadena completa es más importante que cualquier otra cosa.

El concepto nuevo es que muchas de las cadenas de caracteres están relacionadas entre sí. Pueden hacerlo porque están en relación a su origen o al desarrollo histórico o por algunas otras cuestiones. Por ejemplo, entre el latín, el griego, el cirílico y el armenio hay un origen común y tienen muchas formas en común. También hay formas que comparten. Ya las hemos visto en algunos ejemplos en latín y en griego.

En cuanto al japonés, chino y coreano, vemos que hay una superposición de algunas cuestiones y hay algunas características topográficas también que son similares. Lo mismo sucede en India y en países similares que tienen un origen en común, una estructura similar y también formas en común.

Tenemos una investigación en la cual ciertas formas como por ejemplo la Segway UI muestra una tendencia a la utilización de glifos armonizados en todas las cadenas de caracteres. A menudo el griego

se presenta con un estilo un tanto diferente o es posible ver algunas formas en armenio que tienen un cierto cambio. Como estas formas están armonizadas, muchas veces el armenio o el griego se parecen al latín, lo cual hace que se incremente la confusión o la posibilidad de confusión en estas cadenas de caracteres.

Para darles una idea, dos cadenas de Kannada y Telugu básicamente tienen las mismas formas. Lo mismo ocurre en latín y el cirílico, con 28 formas idénticas, y con Devanari y Gomuki, que tienen 26. Es decir, son superposiciones interesantes. Es posible crear una etiqueta en un código de escritura que pueda ser totalmente indistinguible de otra cadena de caracteres similares.

Hay otras cadenas que pueden estar muy relacionadas como Thai y Lao pero que históricamente han evolucionado y han utilizado distintos tipos de fuentes o de letras, y no es posible entonces que en la etiqueta se pueda determinar qué código se utiliza. En el caso del latín y el griego hay varias formas en mayúsculas que se comparten y presentan confusión si el usuario no tiene control de esto. Nosotros tampoco podemos tener control de esto. Siguiente diapositiva.

Algo muy importante a tener en cuenta es que la cantidad de variables definidas para una etiqueta, ya sea bloqueada o no, puede ser muy extensa. Hemos visto ejemplos de cadenas en árabe que pueden tener una gran cantidad de variantes bloqueadas, incluso hasta un millón. Cuando Sarmad habló de crear una lista de variantes bloqueadas, esto es algo que a nivel general o incluso con un ejemplo particular no es posible. Por lo tanto, se requiere otro tipo de técnica porque la

enumeración es un enfoque de fuerza bruta, por así decirlo, que no va a funcionar para estos ejemplos de etiquetas.

Simplemente no es posible comparar todas las variables individuales con otras variables individuales. Hay que utilizar algo que sea diferente, otro método. Probablemente se requiera un proceso de dos etapas en el cual se pueda encontrar una forma automática de eliminar de las comparaciones todos esos pares de etiquetas que probablemente no causen confusión porque contienen características lo suficientemente distintas para que no se tornen confusos. Luego tendremos una lista de etiquetas que van a estar sujetas a revisión y eso se hará a nivel individual.

El punto clave en relación a las etiquetas es tener algún tipo de proceso automático que funcione de manera similar para detectar colisiones en las variantes que están bloqueadas. Actualmente pensamos que un enfoque exitoso podría ser hacer algo similar a lo que se hace con el mecanismo de variantes. Para variantes que están bloqueadas o no, es decir, un sí o un no, y remplazar eso por algo que defina, por ejemplo, el nivel de confusabilidad y que trate este conjunto de variables que prestan a confusión como un conjunto general.

Dado que este conjunto de variables no es transitivo, de alguna manera se supondría que estas variables confundibles son transitivas y se puede ejecutar una herramienta teniendo en cuenta todos estos datos para poder analizar un superconjunto y determinar qué podría ser confundible o no. Podemos pensarlo como una definición de

variantes amplia, de aquellas que no se deberían considerar confundibles y cuáles se podrían excluir. Para el latín, griego y cirílico y armenio tenemos un único punto de código y algunos de los registros tienen políticas que aplican para poder identificar estos puntos de código que dan lugar a confusión.

Hay un conjunto de datos también en Unicode que puede ser confundible. El problema es que no tiene un mantenimiento. Es revisado por el consorcio de Unicode y se requiere alinear esto con la definición de variantes. Por el momento existen bastantes diferencias. También hay muy pocos datos disponibles de las secuencias. Es decir, hay múltiples puntos de código que se confunden con otras secuencias de puntos de código. Por ejemplo, hay cadenas de caracteres que se escriben en clusters. Hay que comparar cluster contra cluster o secuencia contra secuencia.

Incluso en ASCII, si uno escribe .corn en comparación a .com, se puede observar cierto nivel de confusión. Incluso si se hace una evaluación previa hay que determinar si algo se puede confundir o no. Si hablamos del desarrollo de pautas, tenemos que dejar en claro que esto no puede ser un algoritmo. Tiene que existir una decisión en manos del panel de revisión de similitudes. Probablemente tengamos que consultar a los expertos en los códigos de escritura para saber qué es lo que se necesita al momento de analizar las etiquetas. Estamos también pensando en establecer reglas para el uso de los grupos focales porque es un proceso muy costoso y no hay una buena metodología implementada.

Finalmente también reconocemos la necesidad de tener un proceso de preevaluación. En este caso, el panel de revisión de similitud de cadenas de caracteres tendrá que desarrollar un conjunto contencioso teniendo en cuenta lo que Sarmad dijo anteriormente. Actualmente estamos en el medio de la redacción de un informe. Estamos recabando datos iniciales y esto será seguido de un periodo de comentario público para este informe inicial y si la comunidad cree que la metodología propuesta es aceptable, entonces comenzaremos con el proceso de preevaluación de algunas variantes o etiquetas y también etiquetas delegadas para determinar si presentan confusión y luego vamos a redefinir los conjuntos de datos para determinar si son adecuados o no, y ver cuáles son los resultados para finalmente poder terminar con la metodología. ¿Quiere que pasemos a las preguntas?

SARMAD HUSSAIN: Nos hemos quedado sin tiempo para preguntas. Quizá podemos aceptar una o dos preguntas rápidamente. Si no, también hay una dirección de email en el chat que colocó allí Pitinan. Pueden enviar un email a esa dirección y les responderemos. Vamos a darle la palabra a Anil. Hay un micrófono a la derecha, en caso de que quiera tomar la palabra. No podemos escucharlo.

ANIL KUMAR JAIN: Mi pregunta es que supongamos que el panel de revisión de la similitud de cadenas de caracteres ve un dominio que se solicitó y lo aprueba. Luego empieza a tener un verdadero uso en el sistema y en

ese caso empezamos a ver que es similar a un código de escritura existente o que hay reclamos sobre ese código de escritura. ¿Cuál es el procedimiento que se debe aplicar para dar un paso atrás o para tomar medidas? ¿Cuáles son los sistemas disponibles? Gracias.

SARMAD HUSSAIN: Eso probablemente no va a ocurrir, primero que nada. En última instancia, el panel de revisión de la similitud de cadenas de caracteres tiene un sistema de frenos y contrapesos. Hay un conjunto de procedimientos de objeciones que pueden ser utilizados por la comunidad más allá del panel para abordar los problemas que pueden parecer aparentes pero que por alguna razón no son capturados por el panel. Hay frenos y contrapesos entonces, controles en el proceso que también se pueden utilizar. Una vez que ha sido delegado, creo que simplemente se avanza hacia delante.

ANIL KUMAR JAIN: Gracias.

SARMAD HUSSAIN: Vamos a tomar un comentario más y tenemos que cerrar.

NABIL BENAMAR: Por lo que sé, para el código de escritura de árabe esto ya lo hemos hecho. Fue una variante pero para la similitud o para la confusión, esto es algo nuevo. Dentro del mismo código de escritura podemos tener

dos palabras que son similares pero que se escriben en dos idiomas diferentes. Por eso quisiera preguntar si en este panel tenemos un subgrupo que se ocupe del código de escritura árabe, porque tiene algunas especificidades que no encontramos en otros scripts.

SARMAD HUSSAIN: Vamos a desarrollar algunos conjuntos de datos más allá de las variantes que se generaron para considerar las similitudes. Hay algunos paneles de LGR que ya lo han hecho. Cuando esos conjuntos de datos sean necesarios vamos a trabajar con la comunidad para generarlos. Lamento que no hayamos dejado más tiempo para preguntas y respuestas pero pueden ponerse en contacto con nosotros a través del email. Quiero ahora sí agradecerles a todos por venir a esta sesión. Gracias a los presentadores. Se cierra aquí la sesión. Gracias.

[FIN DE LA TRANSCRIPCIÓN]