ICANN82 | Foro de la comunidad – Evaluación de Similitud de Cadena de Caracteres y el Programa de Nuevos gTLD: Próxima Ronda
Miércoles, 12 de marzo de 2025 – 09:00 a 10:00 PST

PITINAN KOOARMORNPATANA: Buenos días. Ya es la hora. Deberíamos empezar. Por favor, comencemos con la grabación.

SARMAD HUSSAIN: Ya comenzamos con la grabación. Adelante.

PITINAN KOOARMORNPATANA: Hola. Bienvenidos a la sesión sobre evaluación de similitud de cadenas de caracteres. Soy Pitinan Kooarmornpatana. Seré la coordinadora de esta sesión. Tengan en cuenta que esta sesión está siendo grabada y se rige por los estándares de comportamiento esperado y la política antiacoso de la comunidad de la ICANN.

Durante la sesión, solo se leerán en voz alta preguntas y comentarios si se presentan en el espacio de preguntas y respuestas. Los comentarios que se escriban en el chat se considerarán parte del chat y quizá no se lean en voz alta. La interpretación para esta sesión incluye inglés, francés y español.

Si desean tomar la palabra durante la sesión, por favor, levanten la mano en Zoom. Cuando lo llamen por su nombre, los participantes virtuales tendrán la posibilidad de habilitar el micrófono en Zoom.

Los participantes presenciales utilizarán el micrófono de la sala para hablar. Por favor, para los registros, digan su nombre y el idioma en el que van a hablar si no lo hacen en inglés. Hablen a una velocidad razonable. Ahora le voy a dar la palabra a Sarmad Hussain. Sarmad, le doy la palabra.

SARMAD HUSSAIN Muchas gracias, Pitinan. Bienvenidos a todos a esta sesión. Buenos días, buenas tardes y buenas noches, según donde estén. Para que tengan una idea general de los temas que vamos a cubrir hoy, quiero decirles que voy a comenzar haciendo una presentación con una descripción breve de la similitud de cadenas de caracteres y el alcance de la evaluación de la similitud.

Luego le voy a pedir a Michael Bauland que nos ayude a describir lo que estamos haciendo en este momento para recopilar los datos sobre similitud de cadenas de caracteres a partir de especialistas en código de escritura. Posteriormente Julien Bernard y David, del equipo de la ICANN, se referirán a la herramienta que se está desarrollando como forma de ayudar a la evaluación manual.

Finalmente veremos cuáles son los plazos que tenemos por delante. Esperamos tener tiempo para preguntas y respuestas al final pero, por favor, no esperen hasta la parte de preguntas y respuestas. Si tienen alguna pregunta en cualquier momento, por favor, levanten la mano en Zoom o escriban su pregunta en la ventana de preguntas

y respuestas o en el chat. También, por supuesto, pueden levantar la mano. Habiendo dicho esto vamos a comenzar.

Para darles una idea general de la próxima ronda de gTLD, de hecho, habrá distintos impactos sobre las cadenas de caracteres. Por un lado habrá que solicitar cadenas de caracteres o variantes que no deben ser similares a cadenas de caracteres que ya existen o que podrían existir.

Obviamente, la motivación subyacente es asegurarnos de que los usuarios finales no se confundan con esas similitudes. El alcance de esta similitud es puramente visual. La similitud obviamente también puede estar al escuchar y en el significado pero en esta evaluación de similitud de cadenas de caracteres está limitada estrictamente a la similitud visual.

Inicialmente hay un par de afirmaciones en SubPro que indican que la similitud de cadenas de caracteres para las cadenas solicitadas debe tenerse en cuenta considerando la ronda de 2012. Además, hay información mucho más detallada en cuanto al alcance y a cómo se debe implementar este proceso en el informe final de fase uno del EPDP que fue un proceso de desarrollo de políticas de SubPro. Obviamente, queremos invitarlos a leer estas recomendaciones específicas que se resumen aquí en pantalla. Queremos invitarlos a que lean exactamente el texto correcto que está en la política.

Para que tengan una reseña de la situación, antes de entrar en los detalles específicos, que son los que figuran en pantalla, creo que vale la pena distinguir la diferencia entre lo que llamamos variantes

y lo que llamamos cadenas similares. En cierta forma, las variantes se definen como aquellas cadenas de caracteres que se consideran iguales en las comunidades específicas que utilizan esos códigos de escritura.

Hay una definición que requiere que las dos cadenas de caracteres no sean diferenciables para la comunidad. Esto incluye, por ejemplo, homógrafos u otras cadenas de caracteres que son similares, no solamente visualmente sino también de alguna otra forma.

La definición de variante se refiere estrictamente al conjunto de reglas para la generación de etiquetas para la zona raíz. Más allá de lo que ya se definió como variantes podría haber algunas cadenas de caracteres adicionales que podrían ser similares aunque no iguales. Por lo tanto, no se definen como variantes.

De todas formas, podrían ser confundibles en caso de que ambas sean delegadas. Podrían ser confundibles para los usuarios finales. Dentro de la evaluación de la similitud de cadenas de caracteres lo que buscamos son aquellos conjuntos de cadenas de caracteres que no son variantes pero que sí se pueden confundir unas con otras. Ese es el alcance de la evaluación de la similitud de cadenas de caracteres.

En función de la recomendación que está en el IDN EPDP se establece el alcance de la comparación. En la parte superior pueden ver cuál es la cadena que se solicita. Podría ser la cadena de gTLD primaria, que alguien la haya solicitado. Esa cadena primaria, ya sea

en ASCII o en IDN puede ser una cadena de variantes asignables y también pueden ser bloqueadas.

Como pueden ver, los tres conjuntos, la primaria, la variante asignable y la variante bloqueada, están dentro del alcance de la evaluación de similitud de cadenas de caracteres. La pregunta es con qué vamos a comparar esas cadenas. Si se fijan, en la izquierda de la pantalla verán todas las categorías que se utilizarán para comparar estas cadenas de caracteres y determinar la similitud visual.

Esto incluye los gTLD existentes, los gTLD solicitados en las rondas anteriores, pero que aún se están procesando. Los ccTLD existentes, los solicitados y otros gTLD solicitados en la ronda actual, los nombres bloqueados y también hay un par de subcategorías de nombres bloqueados que se aplican y luego cualquier cadena de caracteres en ASCII con dos caracteres que se resuelven para los ccTLD.

Todas estas cadenas de caracteres son las que vamos a comparar y no solamente como cadenas primarias sino también todas las variantes asignables y todas las variantes bloqueadas. Por lo tanto, esta será una extensa serie de comparaciones entre las cadenas solicitadas y todas estas otras cadenas y todas las variantes.

La única comparación que está fuera del alcance de este trabajo es ese recuadro en negro que está en la parte inferior derecha, que son las cadenas de caracteres de variantes bloqueadas y las cadenas de

caracteres de variantes que están bloqueadas pero que pertenecen a estas otras categorías que mencioné.

Si se fijan en los casilleros que están en gris, son casilleros que se refieren a comparaciones que el panel de evaluación de similitud de cadenas de caracteres podría utilizar o podría elegir no hacer si no hay posibilidad de hacer una evaluación de similitud de cadenas de caracteres en caso de que pueda haber confusión por cadena de caracteres. Todo está dentro del alcance excepto las bloqueadas comparadas con las bloqueadas.

Publicamos una guía inicial para llevar a cabo la revisión de similitud de cadenas de caracteres. Esto se publicó y se presentó para comentarios públicos hace unos meses. Esa fue una versión inicial basada en nuestro trabajo inicial y en los aportes que recibimos mediante los comentarios públicos. Luego publicaremos una nueva versión de esta guía y la presentaremos para comentarios públicos más adelante, en el transcurso de este año. Estén atentos. Esta orientación es para guiar al panel con respecto a cómo hacer la evaluación de la similitud de cadenas de caracteres.

Como probablemente habrán visto, en la última hoja vamos a tener muchas más revisiones porque no son simplemente cadenas de caracteres comparadas entre sí o con los gTLD o los ccTLD existentes sino también etiquetas variantes de aquello con lo que se está comparando. Habrá muchísimas más comparaciones. Por lo tanto, para ayudar al panel de evaluación de similitud de cadenas de caracteres a hacer su trabajo vamos a desarrollar también una

herramienta de preevaluación. Esta herramienta utilizará datos que están desarrollando los especialistas ahora. Nos vamos a referir a esto en la siguiente sección. La herramienta utiliza estos datos para crear una estimación inicial acerca de lo que podría ser o no similar como para tener una comparación previa. Luego los datos de la preevaluación con el tiempo serán utilizados por el panel de especialistas, el panel de evaluación de cadenas de caracteres, para tomar una decisión final.

La herramienta de preevaluación es simplemente una ayuda para el panel. El panel, por supuesto, puede no considerar las recomendaciones brindadas por la herramienta de preevaluación. Es simplemente una herramienta que utilizará el panel pero no determina nada.

Obviamente, dedicaremos tiempo hoy también a repasar y revisar lo que ofrece la herramienta. Habiendo dicho todo esto veamos ahora si hay alguna pregunta. De lo contrario, pasaremos a hablar acerca de los datos y todo lo que tiene que ver con la herramienta para la evaluación de la similitud de cadenas de caracteres.

PITINAN KOOARMORNPATANA Hay un comentario.

ORADOR DESCONOCIDO Muchas gracias, Pitinan. Gracias, Sarmad, por una clara definición de la herramienta de evaluación de similitud de cadenas de caracteres. Tal como usted dijo, las comparaciones serán muchas.

¿Tenemos una herramienta informática para hacer esta comparación o solamente el solicitante va a tener que hacer esta comparación en forma manual?

SARMAD HUSSAIN Tal como decía, estamos desarrollando esta herramienta de evaluación. Es una herramienta automatizada que ayudará al panel a seleccionar las cadenas de caracteres que potencialmente podrían ser confundibles, dado que el alcance de este trabajo es tan amplio. Es una herramienta de ayuda, no es la que toma la decisión. El panel es el que finalmente toma la decisión con respecto a las similitudes. Crystal pidió la palabra.

CRYSTAL ONDO Quisiera saber si esto considera singulares y plurales o solamente IDN y variantes.

SARMAD HUSSAIN Esto es diferente de similitud por singular o plural. Ese es un proceso aparte. En cuanto al alcance, esto está muy centrado simplemente en la similitud visual.

CRYSTAL ONDO Por lo tanto esa será una conversación aparte, un debate aparte.

SARMAD HUSSAIN Sí, correcto. Hay un par de personas que pidieron la palabra. Aquí en la sala, Abdalmonem. Luego, perdón, no sé su nombre.

TAPANI TARVAINEN Me presentaré. ¿Tengo la palabra?

SARMAD HUSSAIN Todavía no. Ya le vamos a dar la palabra. Empezamos con Abdalmonem.

ABDALMONEM GALILA Mi pregunta es la siguiente. El seguimiento de lo que usted mencionó y de lo que mencionó Anil, ¿cuándo sería la preevaluación? ¿Es antes de que se presente la solicitud? Si uno de los solicitantes pasa esta evaluación y otro trata de tener una de las cadenas de caracteres que podrían ser similares a otra que ya se aceptó, entonces no pasaría por ese proceso. Quedaría bloqueada. ¿Es así?

SARMAD HUSSAIN La herramienta de preevaluación se utiliza una vez que se presentan las solicitudes. No antes de que se presenten las solicitudes sino durante el proceso. La preevaluación no es para los solicitantes. Es una herramienta que utiliza solamente el panel, que es el que hace la evaluación de la similitud de cadenas de caracteres.

Una vez que se presenta la solicitud pasa por un proceso de evaluación y uno de los pasos de ese proceso de evaluación es la

	evaluación de la similitud de cadenas de caracteres que lleva a cabo un panel independiente. Este panel utiliza esta herramienta de preevaluación para determinar cuáles podrían ser cadenas confundibles. Hace una revisión manual y luego crea un conjunto final de cadenas confundibles que luego pasan al conjunto de solicitudes controvertidas.
ABDALMONEM GALILA	Lo que me confunde es ese término de preevaluación.
SARMAD HUSSAIN	Gracias. Podríamos quizá entonces utilizar un término diferente pero esta preevaluación es para el proceso de evaluación manual que lleva a cabo el panel. Está dentro del contexto de cómo lo usamos. Espero que haya quedado claro. Quizá tengamos que cambiar el término la próxima vez. Gracias.
TAPANI TARVEINEN	Hola. Soy representante del grupo de partes interesadas comerciales. No seguí este tema muy de cerca anteriormente. Con respecto a la similitud visual, me parece que podría ser útil definir un poco más el lenguaje. A veces la gente se puede confundir porque no entiende, por ejemplo, yo no entiendo el código de escritura chino. Para mí es confuso. Lo mismo se puede aplicar para el latino y las otras variantes. Lo visual realmente sucede cuando uno lo ve realmente.

SARMAD HUSSAIN

Bien. En las pautas que publicamos hablamos de este tema. Hay algunas suposiciones también que se hacen en ese documento de pautas. Por ejemplo, una de esas observaciones es por ejemplo si alguien que es un nativo, habla con fluidez un idioma, por ejemplo el chino, esa evaluación no se va a hacer en el contexto de alguien que no tiene esa misma cadena o que no comparte ese mismo código de escritura.

A veces se puede utilizar un mismo código de escritura para varios idiomas y esto puede dar lugar a algunas posibles variaciones y la forma en la cual las personas perciben dos caracteres que son diferentes debido al idioma que utilizan escrito o el idioma que hablan a pesar de utilizar el mismo código de escritura.

Estas diferencias obviamente son algo a considerar por parte del panel. No es que quede fuera de consideración. Por otro lado, la evaluación tiene que hacerse también sobre los caracteres visuales de la cadena en sí.

Si la parte visual es en general clara, incluso si hay caracteres que no se utilizan en un idioma en particular pero visualmente alguien que está familiarizado con la cadena lo puede identificar o puede identificar esa diferencia, esas cuestiones son aspectos que el panel toma en cuenta.

La respuesta a su pregunta es que sí. A veces los idiomas crean ciertas diferencias pero aun así llevamos adelante este análisis lo

más cuidadosamente posible para detectar la similitud o la diferencia visual entre las cadenas. Esta es una evaluación y una decisión que toma el panel. Tenemos a Nabil y después hay otros participantes también que quieren tomar la palabra. Después podemos avanzar con el resto de las presentaciones. Si me permiten, vamos a hacer lo siguiente. Vamos a darle la palabra a Nabil y luego vamos a continuar con las presentaciones y después volvemos a usted. Después podemos continuar. Ahora vamos a finalizar la lista de participantes que toman la palabra.

NABIL BENAMAR Gracias. Yo quería agregar algo sobre la similitud o la confusión. Esto va más allá del consenso de la variante porque hemos definido diferentes variantes para tener la posibilidad de distinguir entre dos caracteres o dos etiquetas.

A veces la confusión surge porque se juntan dos caracteres que son similares o que son idénticas a otra letra distinta o separada. Estas son algunas de las cuestiones que hemos agregado a la definición de confusión. Está relacionado solamente con la comunidad que utiliza esa cadena de caracteres. Esto va a depender de cada cadena de caracteres.

STEVE DELBIANCO Los expertos…

SARMAD HUSSAIN Perdón. Ashley levantó la mano también.

STEVE DELBIANCO ¿El panel de expertos va a poder hacer una evaluación de la similitud, incluso si el software en la etapa de preevaluación no lo señala, si esto no es generado en la herramienta?

SARMAD HUSSAIN Como dije anteriormente, esta es una herramienta que nos presta asistencia. La evaluación de la cadena, el panel, por supuesto, tiene la libertad de ignorar o hacer propuestas a lo que dicen.

STEVE DELBIANCO Entonces es para asistirlos. No es restrictivo. Se van a centrar en las cuestiones visuales.

SARMAD HUSSAIN Sí. Esto es algo visual.

STEVE DELBIANCO Sí, entiendo. Entonces esto se va a dar, por ejemplo, supongamos en una cadena larga de un TLD existente. Eso puede que no sea aparentemente visualmente confuso y similar pero, por ejemplo, si hablamos de la edición de una letra, por ejemplo, o tomamos los plurales y los singulares, a veces, al agregar una sola letra crea confusión ya.

SARMAD HUSSAIN A ver, por lo general no hay una sola respuesta a este tipo de preguntas. Por ejemplo, puedo agregar yo una letra o una palabra en inglés y crear una palabra totalmente diferente. Por ejemplo, en algunos idiomas podemos agregar la letra a y eso hace que sea totalmente distinto si agregamos, por ejemplo, dos aes. A veces, una cadena puede tornarse o no confusa pero eso va a depender de la decisión del panel. También va a depender de qué letra se agregue y si se hace esto dentro del contexto de la cadena de caracteres o no. Si se hace a nivel de un idioma, por supuesto este idioma va a tener cierta influencia.

STEVE DELBIANCO La pregunta es entonces si el panel va a poder decidir si, por ejemplo, la adición de una ese a una cadena que es larga causa confusión visual. ¿Va a poder hacer esa conclusión el panel?

SARMAD HUSSAIN Sí, si el panel considera que algo es confuso a nivel visual, entonces sí va a poder tomar esa decisión. Vamos a avanzar. ¿Tienen alguna pregunta?

ASHLEY ROBERTS Con respecto a la evaluación y a la herramienta, ¿esto va a estar disponible para los solicitantes también?

SARMAD HUSSAIN Todavía no. Está en la solicitud pero estamos analizando la herramienta y, con respecto a su pregunta, todavía no hemos terminado de desarrollarla. Una vez que lo hagamos seguramente podamos analizar la posibilidad de que esté disponible en código abierto y a los solicitantes. Todavía no está disponible online para los solicitantes. Por el momento esto está diseñado para el panel pero podemos después lanzar la herramienta para que la gente la pueda utilizar e instalar. Vamos a continuar. Le voy a pedir a Michael Bauland que nos cuente sobre la recopilación de datos en esta evaluación de similitud de cadenas de caracteres.

MICHAEL BAULAND Como dijo Sarmad, actualmente me encuentro gestionando la parte de recopilación de datos. Los datos están definidos por la zona raíz y por las reglas de generación de etiquetas de la zona raíz, y los puntos de código que se encuentran en el repertorio y que básicamente permiten que los TLD, y que estos repertorios que se encuentran dentro de 25 o 26 cadenas, porque se acaba de agregar el Thaana a la raíz, todos estos repertorios tienen que ser analizados.

Esto implica muchos datos a analizar por una sola persona o por un equipo reducido. Esto requiere que el equipo tenga conocimiento de todas las cadenas de caracteres. Por esa razón la tarea está siendo delegada a expertos en cada una de las cadenas de caracteres.

Cada uno de estos expertos tiene un ámbito de trabajar. Tiene que comparar sus repertorios con caracteres en ASCII. Es decir, letras de la A a la Z, tanto en mayúscula como en minúscula. Esto se compara

con el repertorio. También hay una comparación dentro de las cadenas de caracteres. Se comparan los diferentes puntos de código entre sí y también hay una comparación que se hace entre otras cadenas de caracteres porque algunas cadenas de caracteres pertenecen a la misma familia y, por lo tanto, puede haber cierta similitudes y esto requiere que los expertos también puedan comparar sus repertorios con los repertorios de otras cadenas de la misma familia.

También tienen que hacer otras comparaciones. A modo de ejemplo, pueden comparar los elementos en mayúsculas, los caracteres compuestos o los caracteres de forma simple, como una L, los que tienen forma de círculo, etc. o forma de curva.

También las versiones subrayadas. Por ejemplo, si observamos un nombre de dominio en una página web y hay un enlace, esos enlaces suelen estar resaltados y a veces este resaltado oculta partes de un carácter que es muy similar, por ejemplo, a otro que no está subrayado.

Se pidió a los expertos que categorizaran esta comparación y la dividieran en cinco grupos o categorías. Unos son los pares idénticos o casi idénticos, y esto incluye todas las variantes definidas dentro de la LRG de la zona raíz. Aquellas que son muy confundibles o altamente confundibles. Es decir, que son casi homógrafos o muy similares.

Luego tenemos la categoría 3, que son los pares similares, y la categoría 4 y 5 que no son analizadas de forma explícita pero en este

caso, en la categoría 5, aparecen la mayoría de los pares que son básicamente distintos. La categoría 4 incluye pares que son apenas similares. Esto incluye pares que han sido analizados y que se decidió que no pertenecen a las categorías 1, 2 o 3.

Aquí les muestro algunos ejemplos y detalles de las categorías y también verán alguno de los ejemplos que han creado los expertos de las cadenas. Por supuesto, no les voy a mostrar las 25 cadenas de caracteres pero sí hice una selección de algunas que son bastante diversas y que son ejemplos de los datos que hemos recabado hasta el momento.

Con respecto a la categoría 1, es decir, idéntico, casi idéntico más una variante, en este caso se incluye a todas las variantes, automáticamente todas las variantes pertenecen a esta categoría. En este caso no hay que agregarlas manualmente. Esto también incluye otros casos de caracteres idénticos. Por ejemplo, porque están en mayúsculas y quedan fuera del alcance de la variante. También los que son casi idénticos se pueden confundir con facilidad si es que se ven estos caracteres uno al lado del otro. Siguiente diapositiva.

Aquí vemos algunos ejemplos. Por ejemplo, hebreo, latín y bengalí. No los voy a leer porque no sé cómo pronunciarlos pero, como pueden ver, en el caso del latín, si los miramos de esta manera es fácil distinguirlos. Ahora imaginemos que esto está dentro de un enlace y ese enlace está subrayado. El carácter subrayado en el

ejemplo quedaría totalmente oculto y haría que el primer y segundo ejemplo, la primera y segura cadena, fueran exactamente iguales.

Luego tenemos la categoría 2, muy confundibles. Deben ser confundibles para una gran mayoría de usuarios que presten atención cuando no los tienen uno al lado del otro. Aquí vemos dos ejemplos de etiópico y latín. Solo quería que el público tuviera algunos ejemplos de cadenas de caracteres. Pasamos a la siguiente.

Categoría 3. Son los casos similares. Deben ser confundibles para una simple mayoría de usuarios que prestan atención. La diferencia con respecto al caso anterior es que aquí es una simple mayoría mientras que la categoría 2 debía ser una enorme mayoría. Aquí se incluye también el caso especial del código de escritura latino.

Los remplazos o sustituciones de último recurso. Tenemos también el código latín que incluye ASCII. Hay gente que utiliza caracteres que no están en ASCII y que remplazarían por ejemplo los caracteres que tienen un acento en la versión ASCII en caso de que ese carácter no estuviera disponible en el teclado. Aquí tenemos dos ejemplos de latín y bengalí, que caerían dentro de esta categoría 3. Les voy a dar unos segundos para que miren los ejemplos. Pasamos a la siguiente diapositiva, por favor.

Aquí tenemos la categoría 4, que son apenas similares. Es decir, son algo similares pero no son muy confundibles. Serían confundibles por una minoría de usuarios. Los especialistas no están analizando activamente estos casos pero deben tomar nota de estos casos para que la gente que analiza los datos vea que estos casos se

consideraron y no se olvidaron pero se tomó la decisión de que no son lo suficientemente similares como para estar incluidos en las primeras tres categorías. Ahora tenemos dos ejemplos de hebreo y etiópico. La siguiente, por favor.

Finalmente tenemos la categoría de “Diferente”, que es la gran mayoría de pares. Por ejemplo, el caso de algunas letras que no son confundibles en absoluto. No puse ejemplos aquí porque casi todos los pares de cadenas caen dentro de esta categoría.

¿Cuál es el estado actual? En este proyecto de recopilación de datos ya casi finalizamos la mayoría de los códigos de escritura ya se están preparando para presentarse para comentarios públicos. Tenemos algunos que todavía están atravesando el proceso de revisión interna. Este proceso está en curso. Hay dos códigos que están un poco retrasados. Hay uno que se está preparando. Thaana se agregó recientemente a la zona raíz y luego tenemos uno en que los especialistas todavía están trabajando. Es el caso de Myanmar. Hubo algunos casos, algunas situaciones que generaron demora en el trabajo. ¿Alguna pregunta?

SARMAD HUSSAIN

Gracias, Michael. Vamos a tomar un par de preguntas pero solamente dos preguntas porque tenemos que finalizar con el resto de la preparación.

ORADOR DESCONOCIDO Usted habló acerca del sustituto de último recurso. ¿Eso incluye casos en los que tenemos múltiples caracteres?

MICHAEL BAULAND No. Esto no está incluido aquí. Es puramente similitud visual. Si bien en algunos casos sí tiene sentido utilizar este remplazo o sustitución, en cuanto a la similitud visual no es así.

ABDALMONEM GALILA Si hay una palabra que es parte de otra palabra. Por ejemplo, [inaudible]. GOOGLE, Gooole, ¿cómo se manejaría eso? ¿Eso se consideraría una palabra diferente?

MICHAEL BAULAND La parte de recopilación de datos no es a nivel de etiqueta o de palabra. Simplemente comparamos puntos de código o conjuntos con otros puntos de código. Estos datos solamente comparan estos pares y luego la herramienta de software utiliza estos datos para aplicar esto a etiquetas y cadenas de caracteres reales.

SARMAD HUSSAIN La determinación de si son confundibles o no es un paso aparte. Estos son los datos que se van a utilizar para hacer esa evaluación.

ABDALMONEM GALILA Por qué tratan de validar con números. Hay una función dentro de la base de datos que chequea la similitud. Cuando hay más de un 50 %

implica que hay que chequear. Si encontramos un número que podría confundirse con otro. A eso me refería.

SARMAD HUSSAIN Pasemos entonces a la siguiente parte de la presentación. Le voy a dar la palabra a Julien, quien va a hablar acerca de la herramienta que se está desarrollando.

JULIEN BERNARD: Gracias. Yo voy a hablar acerca de la herramienta de evaluación de SSE. Tenemos la herramienta de SSE que está configurada con los datos a los que se refirió Michael. También está configurada con distintos nombres de valores, TLD, ccTLD, nombres bloqueados y toda clase de nombres con los que queremos comparar las cadenas solicitadas.

Esta herramienta va a recibir una lista de todas las cadenas solicitadas a través del sistema TAMS, que es el sistema de gestión de solicitudes de TLD. Va a generar unos informes basados en esos datos. Luego, los miembros del panel podrán analizar estos informes. Si los miembros del panel consideran que no tienen suficientes datos para tomar una decisión, entonces otra vez la herramienta realizará la evaluación.

Tenemos este subconjunto de cadenas solicitadas. Generamos un informe. Luego los miembros del panel pueden analizar este informe personalizado así como el informe que ofrece el sistema. Los miembros del panel podrán editar el resultado de la herramienta

para agregar o quitar cosas, agregar comentarios y darle el resultado al gerente del panel. Este gerente del panel recopilará todos los informes y creará un informe final.

Este informe final luego se volverá a incorporar a TAMS donde se hará la determinación final. Por ejemplo, si se decide que hay una cadena de caracteres que es similar a una cadena solicitada, ahí habrá una controversia. Se rechazará esta cadena que es similar a la ya solicitada o a un nombre bloqueado, etc.

Estoy fijándome para asegurarme de no haberme olvidado de nada. Sé que hubo una pregunta al respecto. Quiero confirmar que la herramienta es solo una orientación y los miembros del panel tomarán sus propias decisiones con la ayuda de la herramienta o con la ayuda del informe que ofrece la herramienta.

La herramienta es una herramienta a la que se puede acceder a través de icann.org. Los miembros del panel tendrán una cuenta y la herramienta puede hacer un análisis automatizado de similitudes de cadenas de caracteres si es que el objetivo es optimizar la revisión manual.

Los datos de similitud de cadena de caracteres descritos por Michael en la diapositiva anterior se agregarán a la herramienta. La cadena de caracteres solicitada se comparará con la lista de cadenas de caracteres, tal como describió Sarmad. Todas las cadenas de caracteres solicitadas, los gTLD existentes, los ccTLD, todo lo solicitado en las rondas anteriores, ASCII de dos caracteres y

nombres bloqueados. Todo esto será considerado en la herramienta.

Básicamente la herramienta llevará a cabo esta tarea en tres pasos. En primer lugar detectará las etiquetas potencialmente confundibles. Luego hará un cálculo de la distancia visual entre esas etiquetas. Si la etiqueta recibe un puntaje que implica que excede el umbral, entonces esto implica que será incluida en el informe final que generará la herramienta y que es la que recibirá el panel.

Este umbral es un valor configurable que determina la sensibilidad de la comparación. Si tenemos un valor de similitud entre un par de cadenas de caracteres que está por debajo de ese valor umbral, entonces la herramienta va a informar el conjunto de variantes que tiene la mayor similitud en este par de cadenas de caracteres.

Con respecto al resultado tenemos dos clases de informes. El informe del sistema, generado en función de todas las cadenas solicitadas. Este informe se generará utilizando valores umbral predeterminados. Luego tenemos el informe personalizado que se genera sobre la base de las cadenas de un subconjunto de cadenas solicitadas. Los miembros del panel pueden también utilizar un valor umbral personalizado. Estos informes van a estar disponibles en formato CSV o HTML. CSV sirve para que los miembros del panel puedan editar los datos y HTML presenta el informe en un formato que permite mostrarlo sin problemas en cualquier herramienta que utilicen los miembros del panel a diferencia del archivo CSV. La siguiente diapositiva, por favor.

El mecanismo básico para la herramienta es el siguiente. En función de los archivos de datos sobre similitud va a calcular si una cadena solicitada es igual o parecida a otras cadenas solicitadas, a tres existentes, a IDN ccTLD solicitados, ASCII de dos caracteres o nombres bloqueados. Las cadenas de caracteres están agrupadas en un conjunto de solicitudes controvertidas potenciales y este puntaje para cada conjunto de solicitudes controvertidas se calcula entre cada par de variantes o de cadenas de caracteres.

El informe para los miembros del panel, para que puedan hacer la revisión final, contendrá el par de cadenas de caracteres potencialmente similares con un puntaje mínimo que está por debajo de un umbral. Habrá un umbral que estará determinado en función de cada par de cadenas de caracteres y habrá un indicador para calcular este valor de similitud. Vamos a tomar el que tenga el puntaje o el valor más bajo. El indicador para calcular este score de similitud se determinará después del análisis de los datos. Pasamos a la siguiente diapositiva, por favor.

El informe de resultados básicamente contiene esta lista de elementos. Algunos conjuntos controvertidos. Va a tener por ejemplo una ID de similitud, la etiqueta 1, la etiqueta 2, por ejemplo, si no es una cadena solicitud. Luego la fuente y la variante para las etiquetas. La que tiene más similitud en esta comparación.

Después vamos a darle un puntaje y tenemos un puntaje para los niveles de similitud que pueden ser menor o igual a tres. Después tenemos comentarios, etc., que son parte también del informe final.

Este es un ejemplo de cómo sería el informe. En este ejemplo teórico tenemos, por ejemplo, un nombre bloqueado donde, por ejemplo, hay una letra remplazada por otra. La l minúscula está remplazada por la I mayúscula. Con esto culmino mi presentación. ¿Alguna pregunta o comentario?

SARMAD HUSSAIN

No veo manos levantadas de los participantes en la sala. Tampoco en el chat. Le vamos a dar la palabra a David, para que brevemente nos hable de los plazos. Después vamos a volver a preguntar si hay preguntas o comentarios.

DAVID WAGHALTER

Muchas gracias. Soy David Waghalter. Soy gerente de producto en el grupo de IT y de ingeniería dentro de la ICANN. Estamos trabajando en esta herramienta. Esta línea de tiempo está en desarrollo. Actualmente nos encontramos en el desarrollo de la herramienta. Casi la mayor parte de nuestros requisitos ya están completados para esta ronda. Estamos finalizando y mejorando los elementos conforme se va requiriendo y continuamos con el desarrollo hasta el mes de abril.

En abril vamos a tener un UAT interno. A partir de allí vamos a analizar si hay algún requerimiento adicional o alguna cuestión adicional que modificar. Una vez que estos e haya evaluado a nivel interno en el mes de junio vamos a desarrollar los resultados y vamos a avanzar con el desarrollo hasta el verano. El objetivo es

hacer nuestra evaluación de calidad y lanzarlo en el mes de septiembre. No sé si hay alguna pregunta sobre esta línea de tiempo.

SARMAD HUSSAIN Gracias, David. Esto nos trae al final de la presentación. Ahora sí, les doy la palabra a los participantes para que hagan preguntas o comentarios. Gracias.

VINNY ADJIBI Buenos días. Quería preguntar lo siguiente. ¿Los datos utilizados para la evaluación van a estar disponibles?

SARMAD HUSSAIN Le pido que se presente.

VINNY ADJIBI Soy Vinny, un participante de NextGen.

SARMAD HUSSAIN Sí. A medida que avanzamos queremos publicar los datos que recibamos de los expertos y del comentario público. Por supuesto, también sobre la base de los comentarios que recibamos de la comunidad vamos a finalizar con la herramienta. Este comentario público creo que va a salir en las próximas dos o tres semanas, si no me equivoco.

Voy a darles otro minuto para ver si hay preguntas o comentarios. Mientras piensan quiero reiterar que esta herramienta está en

desarrollo. Es algo que está en proceso y en esta oportunidad no podemos garantizar que la herramienta esté disponible para la comunidad. Actualmente está disponible para el panel, para que tome las decisiones correspondientes.

STEVE DELBIANCO

Ayer la BC tuvo una presentación y una reunión aquí. La presentación era de aquellos proveedores que ayudan a los registradores a bloquear diferentes registraciones de nombres de dominio y muchas veces las que están bajo ataque son las que tienen mucha solicitud.

Dijeron que había habido un incremento significativo de la efectividad pero teniendo en cuenta la historia de los registratarios dirían que nada de esto se podría aplicar aquí. La pregunta es si están explorando incorporar la inteligencia artificial en la herramienta. Si ya lo dijeron antes les pido disculpas pero quisiera saber eso.

SARMAD HUSSAIN

La inteligencia artificial no se utiliza en la herramienta. Esto se hace puramente sobre la base de cálculos y los rangos de similitud que han brindado los expertos. En al menos un código sí se brindó información automática y supongo que se van a utilizar puntos de código similares, particularmente para el chino. Aquí los expertos trataban de determinar el nivel de confusión de estos diferentes caracteres y el uso de técnicas automáticas para poder hacer esta

clasificación. En los datos que van a comentario público vamos a comentar esta información.

STEVE DELBIANCO: Yo me refería específicamente a la inteligencia artificial no a lo automático. Teniendo en cuenta que vivimos con un enfoque técnico, hay que aceptar que en septiembre, si lanzamos la herramienta y no tiene conocimiento del valor que puede agregar la inteligencia artificial, va a ser un poco complejo. El resto siempre va a utilizar la inteligencia artificial. Tienen que evaluar esta decisión. Estamos en el 2025, son parte de la ICANN. Tenemos que tener en cuenta esto en el proceso.

SARMAD HUSSAIN: Muchas gracias. Vamos a tener en cuenta su comentario.

ABDALMONEM GALILA: Esto va a depender de alguno de los datos que ya existan, si son big data o si está dentro de una base de datos. Depende de varios puntos. De alguna manera esto es distinto. Si por ejemplo la inteligencia artificial depende de buenos puntos, cosa que dudo, sería bueno. De otra manera, me parece que no va a ser una buena opción utilizarla.

SARMAD HUSSAIN	Gracias. De todas formas vamos a tomar en cuenta este comentario, lo vamos a llevar al grupo. [inaudible], adelante, por favor. Está remotamente.
ORADOR DESCONOCIDO	Hola. Yo quería informarles que en nuestro registro, si ustedes desean registrar un nombre de dominio con IDN con una letra similar o en ASCII y la primera letra O, en armenio es similar. A la vista es similar pero no se puede registrar un nombre de dominio utilizando otro carácter. Por ejemplo, si utilizan una primera letra en ASCII, entonces después no van a poder registrar el nombre de dominio. Simplemente lo quería informar porque quizá esto ayude al registro de los sitios que hacen phishing. Esta cuestión de la confusión por similitud.
SARMAD HUSSAIN	Muchas gracias, [inaudible]. Creo que ya se nos acabó el tiempo. No sé si hay más preguntas o comentarios. No veo manos levantadas. Por lo tanto, quiero agradecerles a todos por su participación. También quiero agradecer a los panelistas por toda la contribución que han realizado. Con esto damos por finalizada esta sesión.

[FIN DE LA TRANSCRIPCIÓN]