Problemática-eduPersonTargetedID

AVISO

Este artículo está en revisión. Las próximas ediciones se enviarán por la lista SIR-USERS.

En este artículo describimos la problemática asociada a este atributo, para el cual el cálculo correcto de sus posibles valores puede influir en el uso actual de la federación SIR y en el uso en SIR2.

Definición(es) de este atributo

La definición *tradicional* de eduPersonTargetedID (se corresponde con `urn:oid:1.3.6.1.4.1.5923.1.1.1.10` o, en clave PAPI, con `ePTI`) ha sido:

> *A single string value of no more than 256 characters that uniquely identifiers a user in an opaque, privacy-preserving fashion. In most cases, the value will be different for a given user for each service provider to which a value is sent, to prevent correlation of activity between service providers.*
>
> *Definición extraída del wiki de InCommon Federation*

Es decir, una cadena de no más de 256 caracteres que identifique unívocamente a un usuario de manera opaca, de modo que preserve la privacidad del mismo. Además, esta cadena será usualmente diferente por cada Proveedor de Servicio (SP) al que enviemos dicho atributo, para evitar que puedan realizarse correlaciones entre distintos SP.

Dicho atributo se utiliza en muchos SP para vinculación de cuentas (account linking). Es decir, podemos verlo como una *clave primaria* para la cuenta local (en el SP) que representa al usuario autenticado. Es, por tanto, vital que el valor tenga **persistencia**, se mantenga inmutable a lo largo del tiempo y no se reutilicen sus valores. Los problemas que pueden originarse si el atributo no tiene estas cualidades son:

- un usuario entra en el SP y tiene un perfil distinto al que ya había generado anteriormente (si el ePTI cambia de valor para un usuario)
- un usuario entra en el SP y se encuentra con un perfil de otro usuario distinto (si el ePTI se reutiliza o no se genera a partir de un dato único y no reutilizable)
- un usuario utiliza varios SP y esto permite trazar sus actividades (si el ePTI no es distinto para cada SP)

Hay que tener en cuenta que muchos SP consumen otros atributos del usuario (`mail`, `displayName`, etc.) que permiten "saltarse" la privacidad obtenida con un atributo opaco y que facilitarían por tanto el seguimiento de un usuario en su uso de diferentes SP.

Sin embargo, este atributo, tradicionalmente implementado con un valor de tipo cadena, ya desde la revisión de eduPerson de 2012 cambiaba su definición:

> *`eduPersonTargetedID` is an abstracted version of the SAML V2.0 Name Identifier format of "`urn:oasis:names:tc:SAML:2.0:nameid-format:persistent`" (see http://www.oasis-open.org/committees/download.php/35711). In SAML, this is an XML construct consisting of a string value inside a `<saml:NameID>` element along with a number of XML attributes, of most significance `NameQualifier` and `SPNameQualifier`, which identify the source and intended audience of the value. It is left to specific profiles to define alternate syntaxes, if any, to the standard XML representation used in SAML.*
>
> *Definición eduPerson 2012.*

En lo que respecta a implementaciones, la más forma más extendida durante mucho tiempo ha sido la de "cadena" de la definición "tradicional", pero se aceptaba también la forma "NameID". Desde un tiempo a esta parte, hay implementaciones que sólo emiten (o aceptan) valores como NameID.

Todo lo anterior se complica con la migración de SIR a SIR2, donde ya existía.

Cálculo de eduPersonTargetedID

Una vez hemos visto las características que ha de tener el atributo, veamos cuál sería la mejor opción para generarlo:

- partimos de un atributo del usuario que sea **único**, **permanente** y **no reutilizable**
  - el `uid` o `netId` serían buenas opciones
  - el `mail` podría ser válido si la institución garantiza que no se reutilizarán en ningún caso, y que no cambiarán a lo largo del tiempo
  - no puede utilizarse el nombre, la afiliación y otros atributos no únicos
- tenemos en cuenta la institución origen del usuario
  - para evitar que dos identificadores idénticos en dos instituciones distintas den lugar al mismo valor

- utilizamos el identificador del SP al que quiere acceder el usuario
    - para que la actividad de un usuario en distintos SP no pueda relacionarse
- generamos un hash "salpimentado" con salt, el cual debería permanecer invariable) con los valores anteriores
    - para que el identificador sea opaco

Con todas estas premisas, podemos obtener un ePTI correcto con una fórmula del siguiente estilo:

```
eduPersonTargetedId = HASH(idUsuario + idDestino + idOrigen + "salt")
```

Determinando el Proveedor de Servicio

En federaciones *Hub&Spoke* como SIR2, es el Hub el que habla con los IdP y hay que tener cuidado con la identificación del SP final.

Cuando llega una petición de autenticación SAML a un IdP conectado a SIR2, se recibirán los siguientes datos:

```
<samlp:AuthnRequest xmlns:samlp="urn:oasis:names:tc:SAML:2.0:protocol"
                    xmlns:saml="urn:oasis:names:tc:SAML:2.0:assertion"
                    ID="_c2995790-f7b0-11e8-aeb6-666564636261"
                    Version="2.0"
                    IssueInstant="2018-12-04T10:38:38Z"
                    Destination="https://sso.rediris.es/saml2/idp/SSOService.php"
                    AssertionConsumerServiceURL="https://sir2.rediris.es/hub/SAML2/sml_sp_acs.php"
                    ProtocolBinding="urn:oasis:names:tc:SAML:2.0:bindings:HTTP-Redirect"
                    >
    <saml:Issuer xmlns:saml="urn:oasis:names:tc:SAML:2.0:assertion">https://sir2.rediris.es/hub/metadata/sml
/saml2/</saml:Issuer>
    <samlp:NameIDPolicy Format="urn:oasis:names:tc:SAML:2.0:nameid-format:transient"
                        AllowCreate="false"
                        />
    <samlp:Scoping>
        <samlp:RequesterID>https://sp-remoto.com/sp/</samlp:RequesterID>
    </samlp:Scoping>
</samlp:AuthnRequest>
```

El valor de **`RequesterID`** (en el atributo **`Scoping`**) hace referencia al SP al que quiere acceder el usuario. Es el que utilizaremos para el cálculo de `eduPersonTargetedID`.

El valor de **`Issuer`** (generalmente aceptado como destino en SAML) corresponde al hub de SIR2.

Podemos tener otro problema con la identificación del SP cuando éste utilice distintas URL como destino. Por ejemplo, un servicio de blogs podría utilizar una URL distinta para cada uno de los blogs albergados:

- https://blog1.example.com
- https://blog2.example.com
- ...

En realidad, el SP es único y deberíamos generar un único **ePTI** para todos los blogs. Así pues, puede ser necesaria una *transformación* de la URL obtenida para garantizar un ePTI único por SP (en **OpenID Connect** se maneja el concepto de **Sector Identifier** para esta problemática).

Permitir la *manipulación* de las URL también nos facilita la obtención de ePTI más apropiados en otros casos. Por ejemplo: la FECYT actúa como proxy de acceso a distintos SP; si generamos un ePTI único para la FECYT, se podría relacionar la actividad del usuario en cada uno de los SP finales.

Uso de `eduPersonTargetedID` por parte de los SP

Algunos SP utilizan el valor del ePTI para identificar al usuario. De ahí la importancia de generar un ePTI único y persistente.

Esta funcionalidad hace que los cambios en la generación del atributo puedan ser problemáticos: si un usuario entra al SP con un ePTI distinto, será un usuario distinto para dicho SP.

Un usuario "normal" puede perder los datos almacenados en el SP si cambiamos su ePTI. Pero también los usuarios "administradores" tendrán problemas: no se les permitirá la gestión del SP.

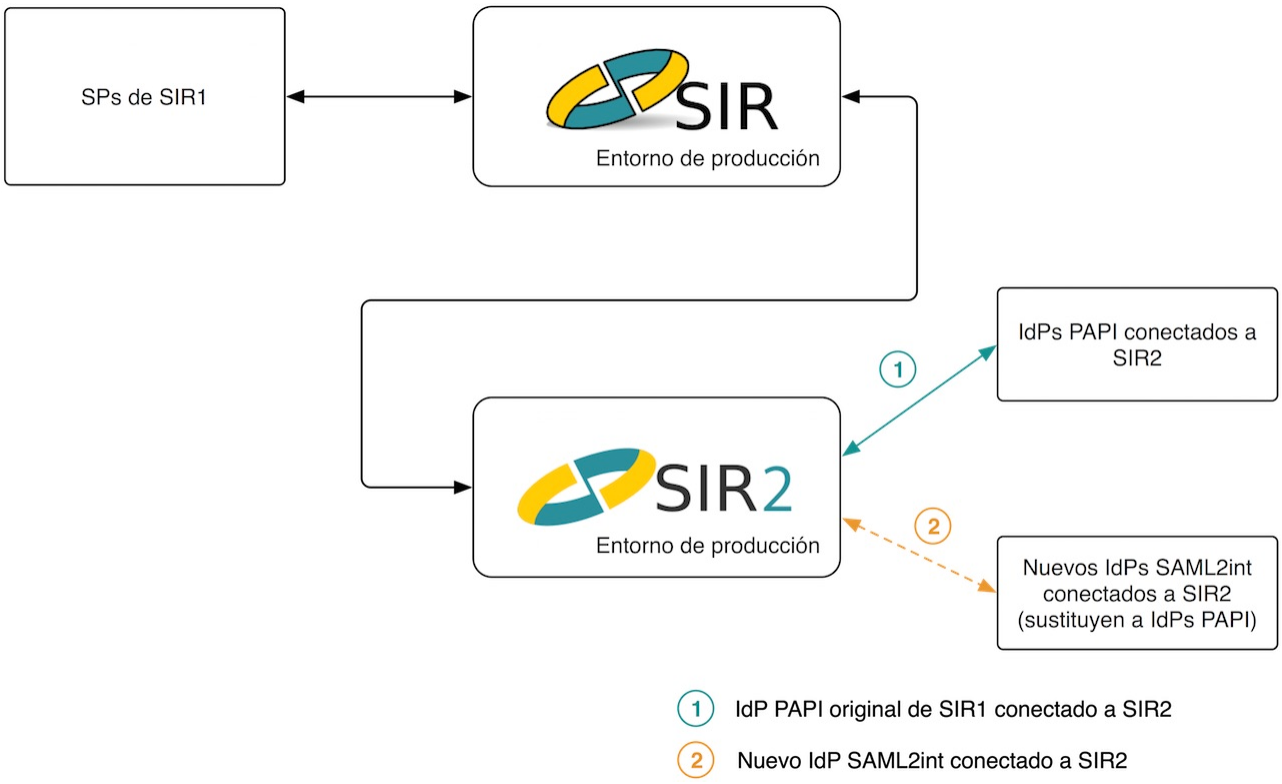
Ejemplo: eBrary utiliza el ePTI para identificar a (todos) los usuarios. Ante un cambio en los valores generados para el ePTI:

- los usuarios estándar perderán las búsquedas almacenadas, las anotaciones, los libros de su estantería, etc.
- los técnicos de la Biblioteca que gestionan la suscripción a eBrary no podrán administrar el portal

Dado que el atributo eduPersonTargetedId puede contener múltiples valores, si el SP hace una gestión correcta del mismo, se podría pasar de un ePTI a uno distinto manteniendo durante un tiempo el valor antiguo y el nuevo valor.

Paso de atributos desde SIR2 a SIR1 durante la migración

Recordamos las conexiones entre sistemas durante la fase 3 de la migración:



Mientras se produce la transición hacia SIR2, ambos hubs van a estar conectados, por lo que tenemos que poner especial cuidado en qué significa cada atributo en cada uno de los hubs, y según el IdP que la institución tenga conectado en un determinado momento. A continuación explicamos por tanto la situación del atributo en ambas federaciones y durante la transición.

`eduPersonTargetedID` en SIR1

En las distintas implementaciones del cálculo del valor del atributo en SIR1, no se ha tenido por lo general en cuenta que el identificador sea *targeted*, implicando esto que se ha reutilizado el mismo valor para todos los proveedores de servicio. Esto, no siendo especialmente grave, hay que tenerlo en cuenta.

Versiones muy antiguas del conector calculaban el valor de `ePTI` (el atributo en clave PAPI) usando el número del proceso Apache, calculado mediante una llamada a la función `getmypid()` de PHP. Esto está lógicamente mal, pues aparte de que se emiten identificadores iguales para distintos usuarios, estos no son persistentes. Este fallo fue comunicado por la lista de correo en su día y la mayoría de proveedores de identidad (IdP) –si no todos– deberían tener el fallo corregido.

Posteriormente se decidió unificar el cálculo de este identificador. En el paso 4 de la Guía de IdPs de SIR1 se decía:

> *Es importante notar que, dado que eduPersonTargetedID es un atributo orientado a preservar la privacidad del usuario y, por tanto, opaco, se ha usado en este ejemplo el hash MD5 del nombre "Antonio...".*

Siendo este el ejemplo, en muchos de los conectores distribuidos por RedIRIS el cálculo se realizaba con la llamada PHP:

```
$assertion = "ePTI=".md5($username."SIR"). ....
```

Que evolucionó a la siguiente fórmula:

```
$epti = sha1($netid);
```

Otras instituciones pueden haber elegido formulas similares a éstas, que **deberán ser revisadas**, sobre todo en función de los proveedores de servicio a los que acceden sus usuarios.

Este atributo ha dado en general pocos problemas y, cuando ocurrieron, estaban relacionados con un cambio de nombre (o incluso de codificación, al añadir acentos o realizar un cambio mínimo en éste).

`eduPersonTargetedID` en SIR2

En SIR2, y desde IdPs SAML2int, tenemos la posibilidad de calcular bien el valor del atributo si no se hacía así hasta ahora, pero tenemos el inconveniente de que si cambiamos los valores de dicho identificador en SPs que vinieran utilizándolo hasta ahora, la vinculación de cuentas anteriores podría perderse.

Se ha desarrollado un módulo para SimpleSAMLphp que permite generar distintos valores para el atributo:

- un valor "correcto", con todas las propiedades requeridas para el ePTI
- un valor "antiguo", que entregue el mismo ePTI que se generaba en SIR1 (para ciertos SP y/o usuarios)

La transición desde SIR1 a SIR2

En SIR1 el valor que se utilizaba para realizar el cálculo de un atributo *targeted*, es el parámetro `PAPIOPOA`. En SIR2 el hub pasa al IdP el valor del `entityId` correspondiente al proveedor de servicio que envió la petición. Estos dos valores se buscará que coincidan en la migración. Es decir, al migrar un proveedor de servicio de SIR1 a SIR2, se mantendrá el mismo identificador, tanto para peticiones PAPI como SAML2int.

A medida que se vayan pasando los SP desde SIR1 a SIR2 se investigará si el SP utiliza el atributo `eduPersonTargetedID` y cómo lo utiliza (si es la clave del usuario, si permite trabajar con múltiples valores, etc.).

Para evitar complicaciones se puede empezar la migración enviando ambos valores del ePTI, para finalizar con un único valor calculado de manera correcta.

Comportamiento del nuevo hub: cacheo de atributos

En la versión inicial del hub de SIR2, se ha tratado de dar prioridad mantener la transparencia en el envío de atributos a SPs que seguirán estando conectados al hub de SIR1, es por ello que el tiempo de duración de las sesiones será corto, prevaleciendo la sesión que se establezca en el hub de SIR1.

~~Posteriormente, a medida que se incorporen SPs directamente al hub de SIR2, tendrá sentido tomar medidas como las siguientes que están en estudio:~~

- ~~**control de sesiones por SP en el hub**. La idea de esta medida es que el hub mantenga sesiones distintas por cada SP, de modo que no compartan valores de atributos, sobre todo de aquellos que sean susceptibles de ser distintos según el SP, tal es el caso de `eduPersonTargetedID`.~~
- ~~**habilitar mecanismos para controlar la duración de una sesión** en el hub por medio del valor de atributos enviados desde el IdP. Tiene cierto sentido que el propio IdP delimite la duración de sesiones en el hub, de modo que para determinados SPs, o incluso usuarios pertenecientes a un grupo, siempre se fuerce la autenticación.~~

Valores en salida de `eduPersonTargetedID`

Otro problema, al que las instituciones deberían ser ajenas, es el de el valor con el que sale el atributo desde SIR hacia el SP.

La implementación

El futuro de `eduPersonTargetedID`

Dada