

WebID como identificador de autores y fuente de información para repositorios institucionales

Trabajo final para “Medida, visibilidad e Impacto de la producción científica”

Alumno: de Albuquerque Pablo

Objetivo

El objetivo de este trabajo es describir mecanismos y protocolos que puedan brindarle a un autor la posibilidad de crear un perfil gestionado por él mismo con la capacidad de exponer sus datos a otros sitios, mediante el uso de Linked Data, utilizando los protocolos WebID y WebID Profile.

Esa información puede ser utilizada por otros sistemas, por ejemplo un repositorio institucional podría crear perfiles de investigadores tomando los datos que crean apropiados para sus necesidades, confiando en la calidad del mismo, ya que es el investigador el responsable de mantener y gestionar su WebID Profile. Como objetivo secundario se permitirá implementar un mecanismo de autenticación single sign-on a través de una clave pública almacenada en el WebID Profile.

Introducción

Una práctica común de los servicios web es centralizar sus recursos, convirtiéndose en lo que se denominan "silos de datos". Cuando nos adherimos a un servicio, como por ejemplo una red social, los usuarios dedican una determinada cantidad de tiempo a describirse a sí mismos completando la información que la aplicación necesita para crear un perfil. De esta forma cada aplicación (o conjunto de aplicaciones que conforman una plataforma para implementar una red social) administra los datos que utiliza, sus mecanismos de identificación y autenticación a gusto. Como resultado, los usuarios tienen dificultades para realizar ciertas acciones como reutilizar sus propios datos en otros sitios o elegir dónde almacenar los mismos. Por ejemplo, al momento de reutilizar la información entregada en otros sitios, el usuario se ve en la necesidad de negociar y adecuarse a los términos que disponga quien posee los datos. Si bien muchos servicios web implementan sus API dedicadas a mejorar la interoperabilidad entre servicios, los desarrolladores deben interiorizarse sobre su funcionamiento y adecuarse a los términos y restricciones que ofrecen dichas API para recién entonces poder implementar aplicaciones multiplataformas (Sambra, 2013).

La realidad descrita anteriormente no escapa al ambiente de los repositorios, donde entidades como los autores son definidas internamente a través del uso de metadatos o mediante el armado de perfiles de autor. Para establecer algún tipo de interoperabilidad entre el repositorio y otros sistemas los autores deben ser identificados y es en este aspecto donde los autores acuden al uso de identificadores controlados por terceros, en el sentido

de que su propiedad, gobierno y descripción están más allá de la jurisdicción de un individuo. Estos identificadores controlados por terceros son ampliamente utilizados ya que existen políticas que prometen su persistencia y procedimientos en caso de que el grupo deje de administrar los identificadores. Por ejemplo, los URI de ORCID, administrados por la comunidad ORCID, permiten identificar un perfil de autor en dicho servicio y son utilizados tanto por investigadores como por organizaciones. Para la mayoría de los investigadores el servicio gratuito ofrecido por ORCID es suficiente para poder describirse a sí mismos, lo que ha generado un alto grado de adopción en la comunidad científica. Una consecuencia involuntaria de este tipo de centralización en la comunicación académica es que un número creciente de instituciones y editores requieren¹ que los investigadores utilicen identificadores ORCID. Los efectos resultantes de tales iniciativas son contrarios a la participación autónoma, en el sentido de que un investigador tiene que ajustarse a los términos y políticas de múltiples terceros. Por otro lado, la administración de la URI por parte de la organización ORCID elimina la carga de mantener la URI y la URL del perfil. En este caso, un tercero "de confianza" se encarga del identificador, mientras que los investigadores son responsables de actualizar el contenido (Capadisli, 2019)

En base al escenario mencionado, uno de los principios² en los que se basa el trabajo que vienen llevando a cabo el NGRWG, es distribuir el control de los recursos y la infraestructura que conforman los repositorios hoy en día, con el objetivo de generar redes distribuidas más sostenibles buscando disminuir, no solo las probabilidades de errores, sino también evitar que un pequeño número de actores puedan tener demasiado control y pueda establecer una posición cuasi monopolista. Una red social distribuida global, pensada en los términos que propone el NGRWG, requiere que cada persona pueda controlar su identidad, pudiendo vincular la misma a través de diversos sitios o sistemas con otros recursos o agentes (personas, organizaciones, etc), ubicando a cada persona en una red de relaciones³. Una de las tecnologías apuntadas para llevar a cabo este desafío es precisamente Web Identity and Discovery (WebID).

WebID⁴ es una propuesta de la W3C realizada en 2014 cuya versión final aún no ha sido publicada oficialmente. WebID ofrece un mecanismo de identificación universal simple, distribuido, extendido que mejora el control sobre la forma en que cada persona puede identificarse para permitir un control de acceso detallado a su información en la Web. En resumen, permite al usuario gestionar su propio WebID. Se trata de un mecanismo distribuido ya que al estar formado por una URI puede ser referenciado desde cualquier lugar de la Web y es extensible porque puede ser utilizado como base para el uso de otros protocolos como por ejemplo un mecanismo de autenticación y control de acceso seguro como WebID-OIDC, WebID-TLS o WebID-TLS+Delegation (Mainini & Laube-Rosenpflanzner, 2016)

Si bien existen entidades como ORCID que ya presentan esta clase de servicios, donde el autor puede publicar sus datos y asociarlos a un identificador, WebID brinda la posibilidad

¹ Requiring ORCID publication workflows open letter -

<https://orcid.org/content/requiring-orcid-publication-workflows-open-letter>

² COAR Next Generation Repositories: Principles - <https://ngr.coar-repositories.org/principles>

³ COAR Next Generation Repositories: Identification of Users -

<https://ngr.coar-repositories.org/behaviour/identification-of-users/>

⁴ WebID 1.0 - <https://dvcs.w3.org/hg/WebID/raw-file/tip/spec/identity-respec.html>

de definir un modelo flexible, que se ajuste a las necesidades del usuario, ya que utiliza ontologías para definir el mismo.

WebID

WebID es un mecanismo extensible, descentralizado y distribuido que permite identificar a un agente (persona, organización, grupo, etc.) a través de una URI dentro de un dominio que idealmente sea propiedad del agente a identificar. Esta URI referencia a un perfil que puede ser interpretado por cualquier usuario web, representado en un archivo bajo el control del agente, llamado **WebID-Profile**, donde se pueden especificar cuestiones como: qué publicaciones posee una persona, a quiénes conoce, cuáles son sus intereses, en que proyectos trabaja, cómo se conforma su curriculum vitae, qué fotos posee, cuáles son sus claves públicas, si utiliza otros identificadores, etc.

El objetivo de WebID es que los agentes creen sus propias identidades, relacionando un identificador único a un perfil que pueda ser interpretado por cualquier usuario web. Para realizar esto, WebID hace uso de serialización en formato RDF. El perfil del agente, contiene toda la información necesaria para crear una red de confianza que permita a las personas construir relaciones entre sus perfiles. Los recursos y relaciones definidos en los perfiles pueden ser utilizados por distintos servicios web.

WebID Profile⁵ es un documento RDF que describe al agente a través del uso de diversos vocabularios como FOAF⁶ y OWL⁷. Este documento puede almacenarse en el propio servidor web del agente, y el acceso al mismo puede estar parcial o totalmente restringido a terceros mediante el uso de controles de acceso, para preservar la privacidad del sujeto del documento de perfil. Dado que el WebID Profile es un documento RDF, su dueño puede utilizar diversas ontologías para modelar la información que crea conveniente y así poder extender su perfil, utilizando los principios propuestos por Linked Data.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<> a foaf:PersonalProfileDocument;
   foaf:maker <#me>;
   foaf:primaryTopic <#me>.

<#me> a foaf:Person;
      foaf:name "Juan";
      foaf:family_name "Gonzalez";
      foaf:mbox <user@example.com>.
```

Ejemplo 1: Un WebID Profile expresado en N3

⁵ WebID Profile or Profile Document -

https://www.w3.org/2005/Incubator/webid/spec/identity/#dfn-profile_document

⁶ FOAF Vocabulary Specification 0.99 - <http://xmlns.com/foaf/spec/>

⁷ OWL - Semantic Web Standards - <https://www.w3.org/OWL/>

En el ejemplo 1 es posible ver cómo se utilizan los vocabularios FOAF y RDF para describir a una persona. En ese documento podemos notar la presencia de dos sujetos que son una URI vacía <>, que significa “este documento” y el recurso <#me>, que hace referencia a la persona que describe el WebID Profile. Del documento (<>) podemos ver cómo se desprenden dos predicados, foaf:maker y foaf:primaryTopic cuyos objetos hacen referencia al recurso <#me> indicando no solamente que este es quien creó el documento sino que también es el objetivo del documento describir a esa persona. Del recurso <#me> podemos saber que es del tipo foaf:Person a través de la propiedad a, (que es un alias para rdf:type). De esta persona sabemos que su nombre "Juan", de apellido "Gonzalez" y cuyo mail es <user@example.com> a través de las propiedades name, family_name y mbox del vocabulario foaf.

Extender el WebID Profile

Una de las ventajas que se pueden mencionar sobre el uso de WebID es la facilidad para extenderlo. Para lograr esto solo es necesario expresar más información a través de otros vocabularios.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix sioc: <http://rdfs.org/sioc/ns#>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.

<> a foaf:PersonalProfileDocument;
    foaf:maker <#me>;
    foaf:primaryTopic <#me>.

<#me> a foaf:Person;
    foaf:name "Juan";
    foaf:family_name "Gonzalez";
    foaf:mbox <user@example.com>.

<https://juan.example.com/blog/posts/1> a sioc:Post;
    sioc:has_creator <#me>;
    dc:title "My first post";
    sioc:content "Lorem ipsum".

<https://myproject.org/> a doap:Project;
    doap:name "My Project";
    doap:developer <#me>.
```

Ejemplo 2: WebID Profile que describe un posteo en un blog y un proyecto utilizando el vocabulario SIOC y DOAP

En el ejemplo 2 es posible ver cómo a través de los vocabulario SIOC y DC, se especifica que la persona “Juan Gonzalez” descrita en el documento, tiene un posteo en su blog, especificando el título del posteo y su contenido.

Existen otros vocabularios interesantes como DOAP que permiten describir los proyectos en los que está involucrada una persona, con la posibilidad de especificar qué rol desempeña en el mismo. Gracias a las ventajas que trae el uso de LinkedData cualquier servicio podría dereferenciar la URI del proyecto (en este caso <https://myproject.org/>) para descubrir más información.

Uso de múltiples identificadores en WebID Profile

Cuando un usuario crea una cuenta en algún servicio o red social, se le asigna un identificador que sirve en el contexto de dicha aplicación, por lo que es normal que un investigador disponga de muchos identificadores (por ej ORCID, Scopus Author ID, WoS ResearcherID, etc.). WebID permite asociar en un perfil otros identificadores que referencien perfiles que puedan dar más información. Una buena práctica es editar los perfiles definidos en otras aplicaciones para que hagan referencia al WebID que el investigador puede controlar. Por ejemplo, a través de la interfaz web de ORCID, un investigador puede modificar su perfil y relacionarlo con otras URIs. De esta forma se crea una relación semántica entre el perfil ORCID y el WebID a través de una propiedad *owl:sameAs* lo que permite que otros programas o agentes informáticos puedan acceder información que no está publicada en ORCID, ya sea por que el modelo no permite incorporar ese dato o simplemente porque no se encuentra disponible.

ORCID ofrece una API a través de la cual es posible acceder a una vista de un perfil en rdf y de esta forma se puede apreciar como la propiedad *owl:sameAs* hace referencia a un WebID gestionado por el propio autor. En el ejemplo 3 podemos ver como en https://pub.orcid.org/experimental_rdf_v1/0000-0002-0880-9125 el autor Sarven Capadisli referencia WebID Profile, del que no solamente es dueño sino que también tiene autoridad a nivel de DNS sobre el dominio *csarven.ca*, donde tiene publicado dicho documento.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

<https://orcid.org/0000-0002-0880-9125>
  a foaf:Person;
  foaf:name "Sarven";
  foaf:family_name "Capadisli";
  owl:sameAs <http://csarven.ca/#i> .
```

Ejemplo 3. Uso de *owl:sameAs* en RDF del perfil de investigador Sarven Capadisli en ORCID

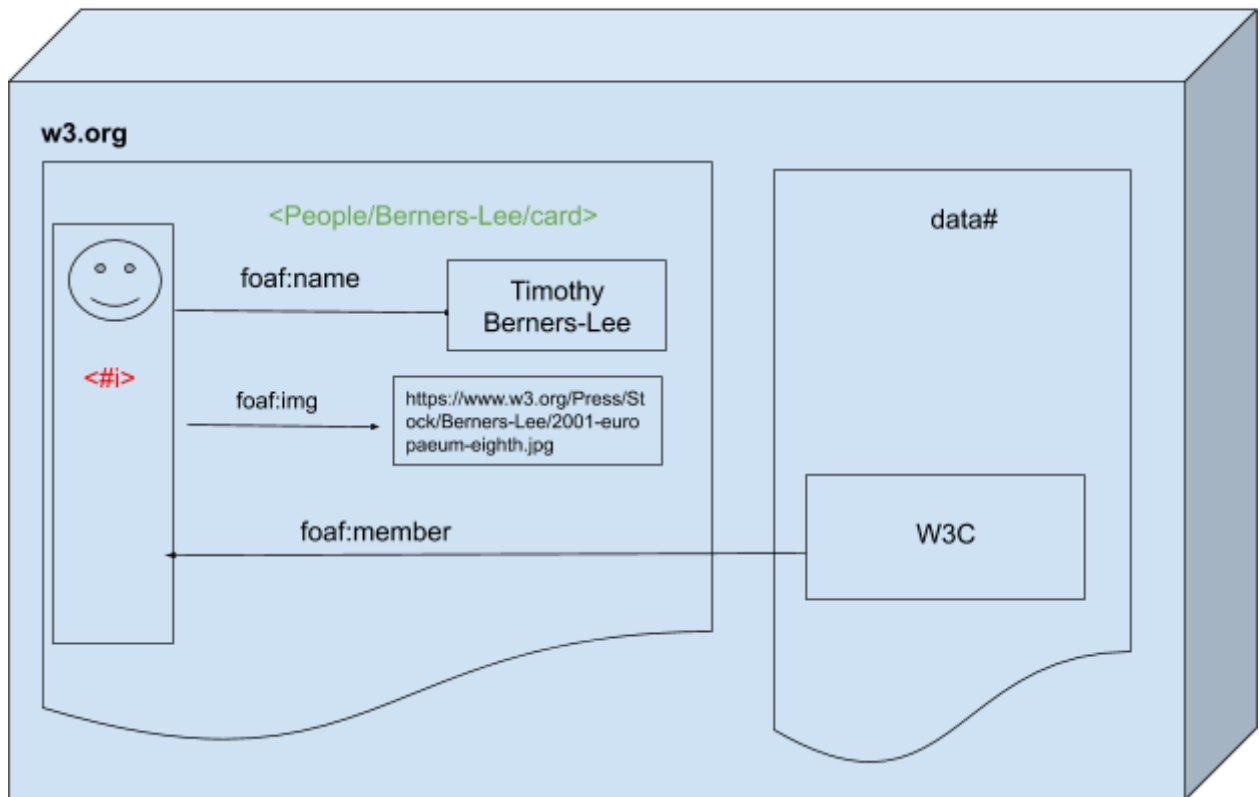
De esta forma las aplicaciones que accedan a la información que brinda ORCID, podrán acceder a través de la propiedad *owl:sameAs* a la información que el investigador publica en su propio WebID Profile, como puede ser una foto o imagen del investigador, dato que no es posible cargar en ORCID.

Uso de WebID en un repositorio institucional

Hemos mencionado anteriormente las ventajas de descentralizar la información que describe a un investigador y de cómo se le pueden brindar herramientas al mismo para tener un mayor control de sus datos, junto con los esfuerzos realizados por grupos como el NGR para incorporar estas prácticas en los repositorios institucionales. Actualmente proyectos como los del software DSpace, utilizado como base para implementar repositorios institucionales, se encuentran adaptando su modelos de datos a estos requerimientos como es en el caso de la versión en desarrollo de DSpace 7⁸. Esta nueva versión adopta un modelo de datos configurable y flexible que permite que una institución pueda representar entidades como autores, proyectos, publicaciones, etc. En el caso del repositorio institucional de la Universidad Nacional de La Plata, implementado en DSpace, parte de la información de sus autores y filiaciones se encuentra en una instancia de Drupal destinada a la gestión de autoridades. Dichas entidades se relacionan a través de identificadores definidos internamente, con los recursos que administra el repositorio (de Albuquerque, 2018). Ante este escenario, el repositorio podría armar un perfil de autor y ofrecer estadísticas de uso basado en las publicaciones gestionadas dentro del repositorio y sumar a la información que se dispone del autor, la definida por el mismo en su propio WebID Profile.

En el ejemplo 4 se ilustra parte de la información que se publica en el WebID profile del investigador Tim Berners Lee, identificado con la URI <https://www.w3.org/People/Berners-Lee/card>. El recurso que hace referencia al investigador es <https://www.w3.org/People/Berners-Lee/card#i> y del mismo se puede obtener el nombre, una imagen y saber que es miembro del W3C.

⁸ Statement from DSpace Steering Group on the DSpace 7 Data Model - <https://duraspace.org/statement-from-dspace-steering-group-on-the-dspace-7-data-model/>



Ejemplo 4. Parte de la información expuesta en el WebID Profile de Tim Berners Lee

Como los datos del ejemplo 4 están siendo modelados a través de la ontología FOAF, es a través de la propiedad *foaf:img* que un repositorio podría acceder a la foto publicada por el autor evitando almacenar y mantener imágenes que podrían no ser el principal interés para un repositorio. La información recuperada del WebID Profile del autor sumada a la que dispone el repositorio, permitiría no solo implementar perfiles de usuarios más completos sino aumentar la visibilidad del autor en otros sistemas que también hagan referencia al mismo recurso.

Conclusión

Si bien WebID es una propuesta que data del año 2014 y su versión final aún no ha sido publicado oficialmente, el hecho de que sea tenido en cuenta por el NGRWG y de que existan implementaciones concretas, hacen de WebID una tecnología a considerar a ser integrada en el desarrollo de servicios pensados para repositorios institucionales. Su uso puede significar no solo un acercamiento hacia la descentralización de la arquitectura de los repositorios, sino también una forma de agrupar varios identificadores en un solo lugar gestionado por el propio investigador y con la posibilidad de extender el modelo de datos ofrecido por terceros, gracias a los beneficios que conlleva el uso de tecnologías relacionadas a Linked Data.

Bibliografía

Capadisli, S. (2020). *Linked Research on the Decentralised Web*.

<http://hdl.handle.net/20.500.11811/8352>

de Albuquerque, P. C. (2018). *Soporte de vocabularios controlados y autoridades en repositorios digitales* [Tesis, Universidad Nacional de La Plata].

<http://sedici.unlp.edu.ar/handle/10915/69754>

Mainini, P., & Laube-Rosenpflanzner, A. (2016). Access Control in Linked Data Using WebID. *arXiv:1611.03019 [cs]*. <http://arxiv.org/abs/1611.03019>

Sambra, A. V. (2013). *Data ownership and interoperability for a decentralized social semantic web* [Phdthesis, Institut National des Télécommunications].

<https://tel.archives-ouvertes.fr/tel-00917965>