UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESINA DE LICENCIATURA

TÍTULO: Desarrollo de servicios basados en perfiles académicos normalizados para autores de repositorios institucionales

AUTORES: Manzur, Ezequiel; Tettamanti, Santiago

DIRECTOR: Dra. Marisa Raquel De Giusti

CODIRECTOR:

ASESOR PROFESIONAL: Lic. Ariel Jorge Lira; Dr. Gonzalo Luián Villarreal

CARRERA: Licenciatura en Sistemas; Licenciatura en Informática

Resumen

Los repositorios institucionales cumplen un rol fundamental a la hora de difundir la producción científica de las instituciones: es por esto que constantemente se busca incorporar servicios que permitan potenciar la difusión de las obras. Esta tesina detalla un profundo análisis y la implementación de un servicio de perfiles de autor para repositorios institucionales, con el objetivo de aplicarlo tanto en SEDICI como en CIC digital. Se desarrolló un prototipo totalmente funcional, que incluye datos reales de los autores de la UNLP. Además, se detalló un proceso para la cosecha y deduplicación de autores desde distintas fuentes (repositorios instituciones, servicios de identificadores persistentes y redes sociales académicas).

Palabras Clave

Identidad digital del investigador; ciencia abierta; perfiles de autores; redes sociales académicas; identificadores persistentes; repositorios institucionales; DSpace; recopilación de datos; normalización de datos

Conclusiones

Se analizaron los modelos de datos planteados por distintos softawares o sistemas que implementan servicios de perfiles de autor y se realizo un relevamiento de los servicios ofrecidos por los repositorios en base a los perfiles de autor. También se llevo a cabo un proceso de recolección, normalización y deduplicación de autores de la UNLP. Esto permitió definir un modelo deseado a implementar y realizar el desarrollo de un prototipo con datos de autores reales de la Universidad Nacional de La Plata.

Trabajos Realizados

Se desarrolló un prototipo funcional que implementa un servicio de perfiles de autor, junto con servicios de valor agregado en torno a ellos, con el uso de datos reales de autores pertenecientes a la UNLP. Para llevar a cabo este prototipo se realizó una extensión de la séptima versión del sofware para repositorios Dspace. También, para poder utilizar datos reales, se realizó una recopilación de datos de autores pertenecientes a la UNLP desde distintas fuentes, los cuales luego se normalizaron y deduplicaron para poder integrarlos al prototipo. Además, se relevaron distintos repositorios y sistemas que implementan el servicio de perfiles de autor para analizar los servicios y modelos de datos que sustentan, con el fin de decidir que servicios y modelo datos utilizar en el prototipo.

Trabajos Futuros

El principal trabajo a futuro que se desprende de esta tesina es el de la implementación real de los servicios desarrollados en el prototipo en los repositorios de SEDICI y CIC Digital. A su vez, quedó pendiente la implementación de varios servicios que podrían haber sido incluidos en el prototipo y que se deberían implementar en un futuro, ya sea en el prototipo o directamente integrarlos a los dos repositorios mencionados. Por otro lado, el proceso de recolección de datos de los autores UNLP podría volver a usarse en un futuro y los datos ser utilizado por una base de datos perteneciente a un repositorio institucional, pero para esto este proceso debe revisar, mejorar y automatizar.

Fecha de la presentación: Junio 2021

Universidad Nacional de La Plata

Facultad de Informática

Tesina de licenciatura

Desarrollo de servicios basados en perfiles académicos normalizados para autores de repositorios institucionales

Autores

Ezequiel Manzur

Santiago Tettamanti

Directora

Dra. Marisa Raquel De Giusti

Asesores profesionales

Lic. Ariel Jorge Lira y Dr. Gonzalo Luján Villarreal

2021

Agradecimientos

Queremos agradecer a nuestras familias por brindarnos su apoyo en todo momento.

A Marisa De Giusti, Gonzalo Villarreal y Ariel Lira, directora y asesores profesionales del trabajo, por su predisposición y dedicación durante la implementación de esta tesina.

Y por último a todo el equipo de PREBI-SEDICI, por su buena voluntad ante cada pedido o consulta durante el desarrollo de este trabajo.

Índice

Índice de figuras

Capítulo 1 - Introducción

Motivación

Impulsados por la constante promoción de las políticas de acceso abierto, la cantidad de repositorios institucionales de acceso abierto y su uso ha crecido considerablemente a lo largo del mundo en los últimos años (Ali *et al.*, 2019). Los repositorios se han convertido en una herramienta clave para la difusión de la producción científica de las instituciones, dado que brindan a los usuarios acceso sin barreras a los recursos científicos y desempeñan un papel importante en el aumento de la visibilidad de los autores (Latif *et al.*, 2018).

Sin embargo, la mayoría de los repositorios carece de un espacio que centralice y exponga la producción académica y científica vinculada a la información personal y/o profesional de sus autores. También es poco frecuente que los autores o la misma institución utilicen los repositorios institucionales como principal vínculo para exponer su producción académica, como por ejemplo en sistemas de currículum en línea o desde redes académicas y científicas. Una de las principales razones por las que el repositorio institucional no se percibe como una herramienta útil por parte de sus autores es el escaso valor añadido que, al parecer, les proporcionan (Barrueco Cruz & Navalón, 2015). En otras palabras, los repositorios institucionales no suelen proporcionar servicios creados en torno a la provisión de nuevos contenidos o bien de contenidos existentes provistos bajo nuevas formas (estadísticas, exportaciones, perfiles de autor, integración con redes sociales) que funcionen como un incentivo para el uso y el depósito de obras en el repositorio por parte de los autores (Genovés, 2017).

En la actualidad, existe una tendencia creciente en el uso de los sistemas RIMS/CRIS (Research Information Management y Current Research Information System, por sus siglas en inglés respectivamente) (por ejemplo, DSpace-CRIS, VIVO), los cuales, en su modelo de datos, incorporan la información de los autores como entidades independientes, ubicándolos de este modo a la altura del resto de

las entidades, como ítems, comunidades o colecciones, permitiendo de este modo la implementación de servicios relacionados a la gestión e interconexión de autores. Asimismo, la comunidad de desarrollo del software DSpace, el software para repositorios más utilizado en el mundo, ha propuesto, para su séptima versión, extender su modelo de datos y funcionalidad para abarcar características de los sistemas CRIS, como el manejo de entidades y la posibilidad de que éstas se relacionen entre sí. De esta manera, se le permitirá a un autor tener sus propios metadatos e interconectarse con otros autores y entidades, como publicaciones, proyectos y organizaciones.

El incremento en la popularidad de los sistemas CRIS, y en especial los cambios introducidos en el modelo de datos de DSpace, abren la posibilidad para que los repositorios institucionales implementen servicios orientados al manejo de perfiles de autor. De esta manera, los autores, al tener un espacio propio en el repositorio, podrán exponer su producción científica y sus datos como investigadores en un solo lugar, que además se adecuará a las políticas de acceso abierto. A su vez, se podrán ofrecer servicios de valor agregado, como el reporte de estadísticas y altmetrics[1], redes de colaboración, interoperabilidad con otros sistemas (por ejemplo, el proyecto de visibilidad web de la UNLP, que interopera con los repositorios para exponer las publicaciones de las instituciones o los autores en los sitios de las instituciones) e integración con identificadores persistentes (por ejemplo, ORCID) (Latif *et al.*, 2018). Todos estos servicios le permiten a un repositorio ampliar el impacto de su producción académica y generar un incentivo más para que los autores depositen sus trabajos en acceso abierto, a la vez que ayudan a mejorar su identidad digital (García Peñalvo, 2017).

El repositorio de la Universidad Nacional de La Plata, el Servicio de Difusión de la Creación Intelectual (SEDICI), contiene la producción científica de gran parte de los investigadores de dicha universidad. SEDICI permite agrupar las obras por autor: posee un buscador donde es posible encontrar a todos los autores del repositorio y permite vincular a los autores con una base de autoridades propia en la que se identifican unívocamente, se normalizan sus nombres y apellidos, y se

[1] Amplio grupo de métricas que permiten medir el impacto de un trabajo o artículo y evaluar a los investigadores, se consideran como alternativas al factor de impacto o los índices de citas de personas usadas para la revistas científicas

validan como miembros de la institución. No obstante, carece de perfiles de autor y de servicios de valor agregado en relación a los autores, no ofrece estadísticas o integración con identificadores persistentes de autor, no se expone la filiación institucional de un autor (centro o instituto en donde trabaja, o su dependencia con otras instituciones) o si tiene perfiles alguna red social académica.

Objetivos

Objetivo general

Maximizar la visibilidad e impacto de la producción científica de los autores de un repositorio institucional y generar un espacio que fomente una mayor interacción entre los autores y el repositorio.

Objetivos específicos

- Analizar la representación de datos de autores en los modelos de datos más usados en el ámbito académico
- Relevar y analizar servicios existentes con respecto a la gestión de autores y sus datos en repositorios institucionales.
- Analizar esquemas de metadatos para descripción de autoridades y evaluar las posibilidades de integración con DSpace.
- Evaluar los mecanismos que ofrece el software DSpace y su comunidad de desarrolladores para gestionar datos de autores dentro de un repositorio digital.
- Diseñar una solución apropiada para el repositorio SEDICI, e implementarla en la herramienta DSpace, a fin de integrar la gestión de autores como un elemento prioritario en el repositorio.
- Incorporar al repositorio mencionado el servicio de perfiles públicos de autores, que incluya sus publicaciones en el repositorio, junto a su dependencia institucional, sus principales áreas de investigación, autores con los que publica frecuentemente, y que brinde accesos a los perfiles de cada autor en redes académicas y científicas.

- Relevar, conseguir y perfilar fuentes de datos de autores existentes en el ámbito interno y externo a la institución.
- Recuperar, normalizar y combinar los conjuntos de datos de autores a fin de construir una base de autores de nuestra Universidad.
- Implementar servicios en torno a los perfiles de autor, que permitan disponer de la información de un autor de diversas formas con el fin de otorgar un valor agregado al repositorio y a su uso .
- Describir los cambios realizados sobre DSpace así como también los métodos aplicados para procesar e importar datos de autores, y sugerir una solución general aplicable en otros repositorios institucionales.

Capítulo 2 - Repositorios institucionales y depósito en acceso abierto

Un repositorio digital institucional es una estructura web que permite organizar, almacenar, preservar y difundir de manera abierta la producción intelectual resultante de la actividad académica e investigativa de una institución; tiene como objetivo poner a disposición de la sociedad la producción científica generada por la institución y hacer que ese contenido sea fácilmente recuperable y disponible (Genovés, 2017). Puede albergar diferentes tipos de objetos digitales, que van desde tesis, trabajos presentados en congresos y artículos de revistas, hasta datos primarios de investigación o documentos institucionales como normativas, ordenanzas, convenios, entre muchos otros.

Son varios los beneficios que aportan los repositorios institucionales a las instituciones, organizaciones y a los investigadores de estas instituciones. Primero y principal, aumentan la visibilidad y el impacto tanto de los autores como de las instituciones al proveer acceso libre y gratuito a la colección de sus publicaciones, lo que proporciona un valor añadido y supone una ventaja competitiva para la institución. Otro beneficio es que en un repositorio abierto el autor conserva sus derechos de propiedad intelectual e integridad sobre la obra.

A pesar de los beneficios, existen algunos obstáculos y reticencias significativas a la hora de incorporar publicaciones científicas en los repositorios. Entre los autores, aunque públicamente la mayoría de ellos apoya el movimiento de Acceso Abierto (AA), según la Comisión Europea en 2018 solamente un 36 % de las publicaciones en revistas y repositorios se encontraban en ese formato (European Commission, 2019; Alonso Arévalo, Subirats Coll & Martínez Conde, 2008). Las razones de que el porcentaje de publicaciones en AA sea tan bajo son variadas, y a menudo los repositorios son poco conocidos o padecen de una falta de visibilidad dentro de su institución. A su vez, muchos investigadores son conservadores respecto del sistema de comunicación académico y no desean cambios fundamentales en la

forma en que su investigación es diseminada y publicada; otros no conocen qué es el AA o sus ventajas e implicaciones, o bien desconocen el sistema de autoarchivo (COAR, 2013). La falta de incentivos por parte de los repositorios es también un factor importante, puesto que la mayoría de los repositorios no poseen servicios que motiven a los autores a depositar sus obras, lo que hace que éstos no perciban que la publicación en los repositorios conlleve algún beneficio para su carrera o prestigio profesional. Otras razones incluyen la preocupación de los autores por infringir la legislación de derechos de autor o de violar los acuerdos de publicación con sus instituciones u organizaciones, o también la falta de sanciones por parte de las instituciones para los autores que no depositen en acceso abierto.

Capítulo 3 - Perfiles de autor

Introducción

Con el aumento de la producción científica se ha tornado cada vez más relevante la desambiguación de los autores (Escolar & Ruiz, 2012). Es común, en los repositorios institucionales, al agrupar las obras por autor, que aparezcan dos autores con el mismo nombre, que un autor sea conocido con distintos nombres (variantes), o incluso puede suceder que bajo el mismo nombre de autor se confundan obras de dos autores distintos (Latif *et al.*, 2018; Ferreira *et al.*, 2013).

Un ejemplo de variante en el nombre de un autor se puede ver en la búsqueda por autor de SEDICI, el repositorio institucional de la UNLP, como se muestra en la figura 1, en donde se muestra (en apariencia) dos autores: «Paracampo, Ariel Hernán» y «Paracampo, Ariel», con una y dos publicaciones, respectivamente. Sin embargo, si se hace clic sobre cualquiera de ellos, se es redireccionado hacia una misma página, que muestra las publicaciones del autor «Paracampo, Ariel Hernán» (ver Figura 2), con un total de tres publicaciones realizadas, lo que deja entrever que estos dos autores son el mismo.

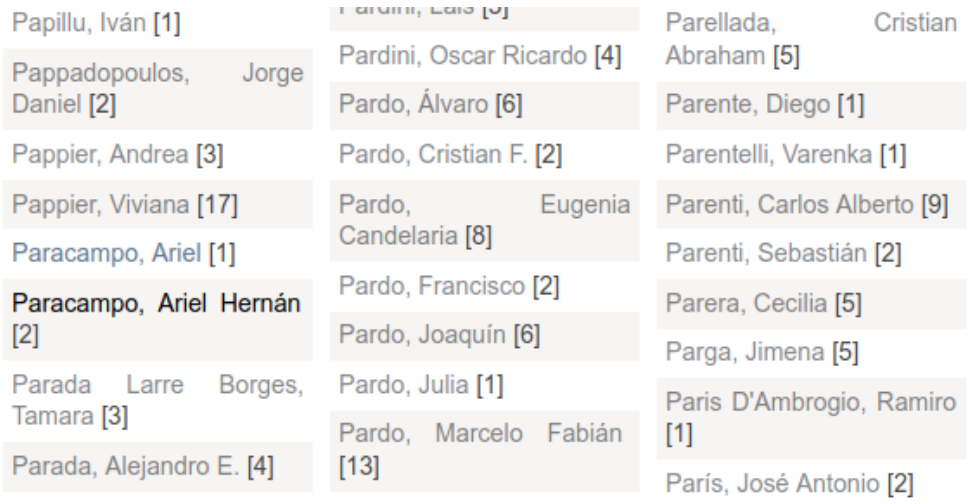


Figura 1. *Ejemplo de búsqueda por autor en SEDICI*

Figura 2. *Búsqueda de las publicaciones del autor «Paracampo, Ariel Hernán» en SEDICI*

Históricamente ha habido diversas iniciativas para mitigar este problema, y en particular se pueden destacar dos de ellas: los identificadores persistentes y los perfiles de autor. Según las definiciones propuestas por Lorenzo Escolar & Pastor Ruiz (2012), los identificadores persistentes son códigos numéricos o alfanuméricos que se asignan a un autor para identificar de forma unívoca su producción científica. La segunda iniciativa, los perfiles de autor, en cambio, se refieren al conjunto de datos que recogen de forma estandarizada bien únicamente la producción científica de un autor o bien toda su actividad investigativa (puestos desempeñados, proyectos de investigación, contratos, etc.).

Los perfiles de autor permiten enriquecer la información de un investigador, al reunir toda su producción disponible en el repositorio, mostrar datos personales y de filiación, estadísticas de productividad y uso y alertas, entre otras funcionalidades; además de proveer enlaces a otros identificadores y perfiles de los autores. Sumado a esto, los perfiles dan la posibilidad a un repositorio de agregar servicios de valor agregado que, por un lado incentivan a los investigadores a incrementar el uso del repositorio, realizar autoarchivo y aumentar las estadísticas de sus publicaciones, y por el otro ayudan a mejorar la visibilidad de la institución a la que pertenece el repositorio (Genovés, 2017). Además de ser un servicio de valor agregado en sí mismos, los perfiles de autor también favorecen la generación de otros servicios de este tipo alrededor de ellos, a partir de los datos que el perfil expone, como los reportes de estadísticas, la exportación de los datos o publicaciones de un autor, redes de colaboración de autores, armado automático de currículums o incluso la exposición de los datos del autor en algún protocolo de interoperabilidad (OAI, Open Search o Web-ID).

Tipología

Según Lorenzo Escolar & Pastor Ruiz (2012), se pueden clasificar los sistemas de perfiles o de identificación de autores según estén relacionados o no con un identificador de autor y si son capaces de integrarse con otros sistemas de comunicación científica. Estos autores plantean la siguiente clasificación:

- *Sistemas de identificación puros*: sólo se centran en el desarrollo de un identificador persistente de autores, por ejemplo ISNI (International Standard Name Identifier, similar al ISBN de los libros pero para autores).
- *Sistemas de perfil puros*: se limitan al desarrollo de un sistema de currículum normalizado, sin asignarle ningún identificador (CVar de Argentina, Lattes de Brasil o CVN de España).
- *Sistemas mixtos*: agrupan los dos anteriores y permiten adjudicar tanto un identificador como un perfil a cada autor (Scopus Author Identifier, ResearcherID).

- Por último, los *sistemas globales*: integran identificadores y perfiles generados por cualquier otro sistema (ORCID).

ISNI se destaca como sistema de identificación puro entre las iniciativas que han intentado imponer un estándar de identificación unívoca de autores. Es un desarrollo de la norma ISO 27729, y fue diseñado para la identificación de personas y organizaciones involucradas en actividades creativas, así como personajes públicos de ambas, como pseudónimos, nombres artísticos y otros. Es un estándar abierto (aunque se debe pagar para obtener el identificador), y ampliamente usado, posee registros de más de 10 millones de individuos[2]. Otro sistema de identificación es Iralis, un sistema de estandarización de las firmas de los autores científicos. Su objetivo es crear un registro de autoridades e intenta producir una base de datos con todas las variantes de firma utilizadas por cada autor, y es de carácter gratuito.

Los sistemas de perfil puro se caracterizan por la recopilación de las obras de un autor y de sus datos personales y de investigador, sin asignarle ningún identificador alfanumérico. Se pueden ver ejemplos de este tipo de sistema en sistemas curriculares on line, como CVar, un registro nacional de los datos curriculares de personal científico argentino, o en varios repositorios institucionales, donde se utilizan para reunir toda la información de las publicaciones de un autor, mostrar su información personal y de investigador. Un caso de uso de los sistemas de perfil puro son los llamados sistemas CRIS o RIMS (Research Information Management Systems). Estos sistemas recopilan y almacenan datos estructurados sobre la investigación del profesorado y las actividades académicas para una institución, con la intención de reutilizar la información de diversas maneras (Keeping up with RIMS, n.d.).

Los sistemas mixtos se caracterizan por ofrecer un código alfanumérico asociado a un perfil, por lo que combinan las características de los dos tipos de perfiles analizados anteriormente. Quizás las dos iniciativas más destacadas en este aspecto sean Scopus Author Identifier y ResearcherID, comerciales ambas. ResearcherID es un producto abierto a todos los investigadores, que pueden darse

[2] Dato obtenido desde el sitio web de ISNI: https://isni.org/

de alta, obtener un identificador persistente y gestionar su perfil que queda asociado a ese identificador. Scopus Author ID, por su parte, es implementado por la editorial Elsevier y asigna automáticamente un identificador y un perfil a todos los autores que se encuentren en su base de datos.

Los sistemas globales responden a un sistema más abierto y completo: permiten incorporar datos de diversas fuentes e interoperar con ellas, e integran identificadores y perfiles elaborados por cualquier operador; esta capacidad de interoperar e integrarse con otros operadores es la principal diferencia con los sistemas de perfiles mixtos. ORCID es quizás el ejemplo más preponderante actualmente de este tipo de sistemas: usado en más de 1200 sistemas (Abril 2021)[3], pretende establecer un registro abierto e independiente asignando identificadores únicos enlazables a la producción científica del autor, independientemente del portal científico en el que aparezca. Es de código abierto y permite la integración con prácticamente cualquier tipo de sistema, conectándose con la información del autor a través del identificador único, lo que posibilita la actualización automática de la información en todos los sistemas integrados (Meadows, 2017).

Visibilidad e impacto

Como se vio en el apartado de repositorios institucionales y el depósito en acceso abierto, sólo un bajo porcentaje de los autores de una institución depositan su trabajo en AA y en repositorios de su institución. Las posibles razones por las que esto sucede pueden ser el desconocimiento por parte del autor tanto del repositorio de su institución (quizás a raíz de su poca visibilidad o por falta de difusión de parte de la institución) como de los mandatos de depósito, o simplemente por no ver algún beneficio personal en el depósito por la presunta falta de servicios de valor agregado. Para mantener los repositorios institucionales relevantes para los autores en un mundo donde son cada vez más populares las redes sociales académicas y sistemas de perfiles globales, es necesario que el repositorio sea visto como una parte integral de la infraestructura de investigación

[3] Dato obtenido desde la web de ORCID: https://orcid.org/members

de la institución y como una herramienta que proporciona mejoras que beneficiarán a los académicos en la promoción de su investigación (Lemberget *et al.*, 2017).

En concordancia con lo planteado por García Peñalvo (2017), se señala que el servicio de perfiles de autor en un repositorio institucional ayuda a percibir a este último como algo más que sólo un lugar de depósito para las obras de los autores. Los perfiles ayudan en la puesta en valor de la investigación de un autor y permiten mostrar su impacto, al ser un medio para su difusión y dar soporte a indicadores alternativos (altmetrics), lo cual revierte en un incremento de las citas. Además, estos indicadores se convierten en un factor de identidad y reconocimiento que luego se traduce en un aumento de la visibilidad del investigador, que, a su vez, ayuda a aumentar la visibilidad general del repositorio y, por extensión, de su institución.

La aparición de una mayor cantidad de servicios de valor agregado es otra de las consecuencias de la implementación de perfiles de autor. A partir de los datos agrupados en el perfil es posible generar una amplia variedad de servicios, como estadísticas, hojas de vida o currículums para investigadores, reporte de cantidad de citas, redes de coautoría, y se puede relacionar al autor con identificadores persistentes, además de interoperar con distintos sistemas de perfiles mixtos o globales como ORCID. Todos estos servicios son muy importantes al momento del incentivo a los investigadores para que depositen sus obras: hay casos en donde se ha encontrado en este método una estrategia exitosa para poblar el repositorio, lo que tiene como resultado un aumento en el uso del repositorio, y, al aumentar la cantidad de obras en él, también repercute en un aumento de su visibilidad (COAR, 2013).

Estado del arte de los perfiles de autor en repositorios institucionales

A pesar de las múltiples ventajas que acarrea el uso de perfiles de autor, estos son poco usados por los repositorios institucionales. Para febrero de 2017, de los

primeros 50 repositorios institucionales mejor puntuados en el Ranking Web de Repositorios de Webometrics, los resultados indican que sólo 5 repositorios (10 %) proveían el servicio de perfiles de autor (Genovés, 2017). Aunque una estadística actual (Noviembre 2020) indica que ese porcentaje aumentó al doble (ahora 10 de los primeros 50 repositorios del ranking los han implementado), el porcentaje sigue siendo bajo si se observan los primeros 200 elementos del ranking, en donde únicamente 28 repositorios utilizan perfiles de autor, es decir, un 14 %. Un análisis más detallado puede verse en el capítulo 5 de esta tesina.

Estos 28 repositorios corresponden a perfiles de autor puros: solo se limitan a recopilar los datos de algunos de sus autores y agrupar sus publicaciones en secciones del repositorio dedicadas a este fin. Sin embargo, varios de estos repositorios tienen alguna característica que los acerca a la clasificación de sistemas mixtos o incluso sistemas globales: muchos permiten la integración con proveedores de identificadores persistentes como Scopus ID, ORCID o ResearcherID, y algunos proveen a los autores con algún identificador persistente propio de la institución a la que pertenece el repositorio, aunque estos identificadores propios de la institución sólo se usan para un control de autoridades interno.

Capítulo 4 - Modelo de datos

Introducción

En el abanico de sistemas web dedicados a la investigación académica existen varios de ellos con implementaciones de perfiles de autor. Aunque, como ya se mencionó, los perfiles de autor no son tan comunes dentro del conjunto de repositorios institucionales, sí se los suele encontrar en los sistemas proveedores de identificadores persistentes, como ORCID, Scopus, o Publons (servicio web proveedor del ResearcherID); en redes sociales académicas, como ResearchGate o dentro de los sistemas RIMS/CRIS (por ejemplo, VIVO). Si bien, con respecto a los perfiles de autor, todos los sistemas que siguen el modelo CERIF disponen de un modelo de datos similar, esto no ocurre con el resto: en su mayoría son iniciativas privadas que han desarrollado sus propias implementaciones y que persiguen objetivos distintos. En consecuencia, a pesar de que se modela una misma entidad (una persona), los modelos de datos en las implementaciones de perfiles de autor son diferentes. Por ejemplo, los proveedores de identificadores persistentes toman al identificador persistente como un dato central, mientras que para una red social académica no es tan importante (ResearchGate ni siquiera cuenta con ese dato); o quizás para un sistema CRIS, el proyecto o laboratorio en donde se encuentra trabajando el autor es un dato relevante; en cambio, a Publons sólo le interesa la institución de la que es parte, sin tener en cuenta la posición del autor en su estructura interna.

A lo largo de este capítulo se realiza un análisis de los distintos modelos de datos en varios de los sistemas que implementan perfiles de autor. Se releva tanto los datos propios de un autor (sus datos personales, sus áreas de interés, biografía, título profesional), como las relaciones con el resto de las entidades (instituciones, proyectos, departamentos, publicaciones, entre otros). Luego, se define un modelo de datos propio, a partir del análisis realizado al resto de los sistemas, con la inclusión de las características más relevantes y que mejor

encajen con el tipo de perfil de autor buscado. Entre los sistemas analizados se encuentran los proveedores de identificadores persistentes ORCID y SCOPUS ID, el modelo CERIF y DSpace CRIS como una implementación de ese modelo, DSpace en su versión 7 con su modelo de entidades dinámicas y un análisis del modelo de DSpace, el cual no tiene al autor como una entidad independiente pero sirve de base tanto para DSpace CRIS y DSpace 7, como para distintas implementaciones de este software en repositorios institucionales que extienden su modelo de datos para poder realizar una mejor gestión de sus autores.

Análisis de modelos existentes

ORCID

ORCID es un sistema abierto para que los investigadores compartan información a escala global. Este se esfuerza en permitir conexiones entre los investigadores, sus contribuciones y sus afiliaciones al proporcionar un identificador único y persistente para que las personas lo utilicen mientras PARTICIPAN en actividades de investigación, becas e innovación (ORCID, 2021).

Para crear un perfil ORCID el autor se debe registrar en el sistema y de esta forma se genera un perfil de autor donde la persona autenticada tiene el control sobre sus datos. También se pueden agregar usuarios de confianza (*trusted individual*) para actualizar ciertos registros del perfil.

En el modelo de datos propuesto por ORCID, se registra información personal (biografía, sitios web personales, país de origen, correos electrónicos, entre otros) e información profesional (educación, empleos, instituciones que financian los proyectos, cargos y distinciones, membresías).

En un análisis más profundo, se puede ver que cada sección de datos se compone de múltiples campos y que permite tener distintas configuraciones de visibilidad.

Configuraciones de visibilidad: Se permite configurar la visibilidad de cada uno de los datos por separado, con tres opciones posibles:

- *Todos*: Toda persona que acceda al perfil público puede ver los datos.
- *Personas de confianza*: Los registros que tengan esta configuración pueden ser actualizados por los usuarios o instituciones de confianza agregados por cada autor.
- *Sólo yo*: La información sólo puede ser vista por el autor. Es importante destacar que esta información es utilizada por el algoritmo de ORCID para desambiguar personas.

Listado de campos y cómo se componen:

- *Nombre y Apellido* (nombre, apellido y nombre publicado)
- *Biografía*
- *También conocido como*
- *País*
- *Palabras clave*
- *Sitios web* (descripción y URL)
- *Otros ID*
- *Correos electrónicos*
- *Educación* (institución, ciudad, estado, país, departamento, rol/título, fecha de inicio y fecha de fin)
- *Empleo* (institución, ciudad, estado, país, departamento, rol/título, fecha de inicio y fecha de fin)
- *Cargos invitados y distinciones* (institución, ciudad, estado, país, departamento, rol/título, fecha de inicio y fecha de fin)
- *Instituciones que financian los proyectos* (tipo, sub-tipo, título, descripción, cantidad, fecha de inicio, fecha de fin, URL y nombre, ciudad, estado y país de la agencia de financiación)
- *Membresía* (institución, ciudad, estado, país, departamento, rol/título, fecha de inicio y fecha de fin)
- *Obras* (categoría, tipo, título, subtítulo, fecha de publicación, país, datos de identificadores de la obra y datos para citar el documento)

DSpace

El modelo de datos de DSpace consta de una estructura de comunidades, colecciones e ítems y metadatos, donde una comunidad puede agrupar varias subcomunidades o colecciones y una colección agrupa varios ítems, como puede verse gráficamente en la figura 3.

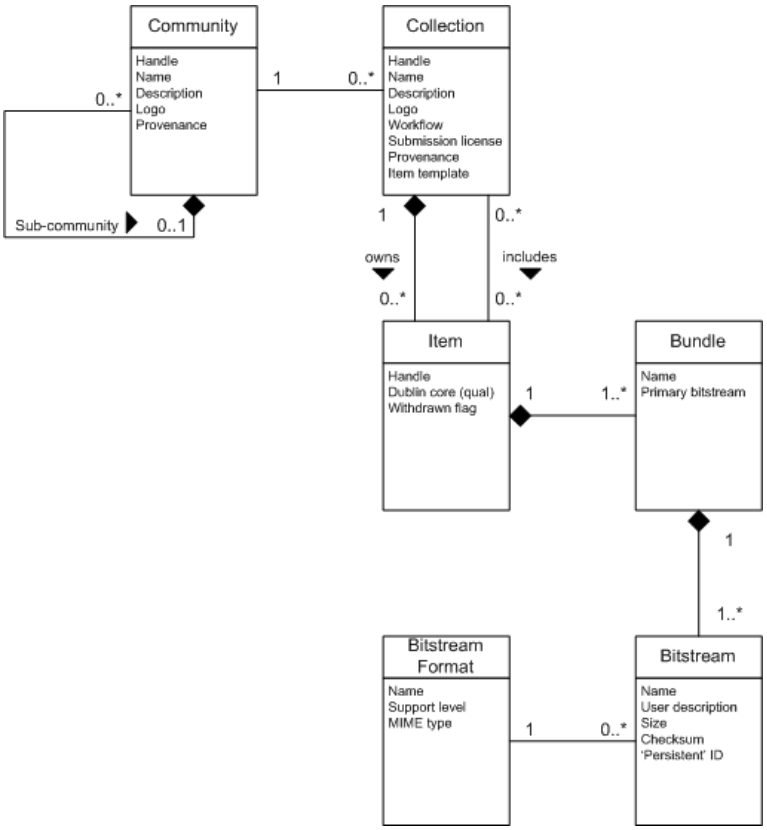


Figura 3. *Modelo de datos de DSpace*

Un ítem pertenece a una colección y puede estar asociado a otras colecciones. Las colecciones solo tienen ítems y pertenecen a una o más comunidades, las cuales, a su vez, pueden contener subcomunidades (DSpace, n.d.).

Los ítems, que representan a las publicaciones del repositorio, son descritos a través de metadatos, que poseen información acerca de los autores, la fecha de publicación, materias y demás elementos que contextualizan una publicación. Los metadatos se almacenan en su propia tabla de la base de datos, «metadatavalue», la cual contiene el ID del ítem al que hace referencia, el nombre del metadato y el ID del esquema de metadatos bajo el que se encuentra (Dublin Core, DCTerms u otros) Esta información se preserva en otra tabla de nombre «metadatafieldregistry», y el valor del metadato en el ítem. Los metadatos pueden

estar relacionados con una clave de autoridad, que conecta el metadato con algún vocabulario controlado y permite normalizar la información de los recursos y conectarlos con fuentes externas, por ejemplo, en SEDICI se puede conectar a un autor de una publicación con la correspondiente servicio web proveedor del ResearcherIDpersona en la base de autoridades del repositorio a través de su clave de autoridad. A su vez, los ítems contienen *bundles* y estos contienen a los *bitstreams*, que representan los archivos u objetos digitales del repositorio.

DSpace-CRIS

DSpace-CRIS es una extensión de código abierto del software DSpace que incorpora las características de los sistemas CRIS/RIMS. DSpace-CRIS amplía la funcionalidad de DSpace y expande su modelo de datos mientras mantiene su base de código.

La principal característica de DSpace-CRIS es su modelo de datos flexible, que le permite recopilar y gestionar datos e información de investigación típicos de un sistema CRIS, para definir entidades y atributos con sus vínculos recíprocos. Dentro de las mencionadas entidades se encuentran las personas, de las cuales, mediante los vínculos y relaciones con otras entidades, se pueden conocer sus publicaciones, organizaciones a las que pertenece, proyectos en los que participó, eventos, colaboradores, etc.; además de datos propios de una persona, como sus nombres y variantes, identificadores persistentes, áreas de interés, profesión y filiaciones, biografía y una imagen personal. Tanto las relaciones como los datos propios de la entidad son configurables y expandibles, con sólo algunos campos predefinidos. También cuenta con útiles funcionalidades como gráficos de red de colaboración, bibliometría y estadísticas agregadas (por investigador, por departamento) con reportes gráficos, CV y bibliografías, integración con ORCID API, entre otros (Lyrasis | DSpace-CRIS Home, n. d.).

DSpace 7

DSpace, en su versión 7, aún en desarrollo, extiende su modelo de datos para soportar entidades de distintos tipos, como personas, organizaciones y proyectos (en las versiones anteriores sólo existen los ítems o publicaciones) así como

relaciones entre ellos.. El modelo tiene similitudes con el Portland Common Data Model (Duraspace, 2018), pero utiliza una estructura de grafos similar a la usada en los sistemas CRIS.

El único cambio realizado al modelo original de DSpace es la inclusión de tres tablas de base de datos separadas en las que se almacenan las relaciones, el tipo de entidades y los datos de una relación, como se puede ver en las figuras 5, 6 y 7. La tabla 'entity_type' solo contiene los nombres de los tipos de entidades (personas, organizaciones, revistas, publicaciones, etc.) con sus ID; la tabla 'relationship' alberga el ID de los ítems que participan en una relación entre entidades junto con el ID del tipo de relación que estos ítems representan; por último, la tabla 'relationship_type' contiene el tipo de entidades que participan en la relación (personas y publicaciones para mostrar autores de una publicación, revistas y volúmenes para volúmenes de revista, personas y organizaciones o personas y proyectos para miembros de una organización o de un proyecto, etc.), junto con datos adicionales, como la cardinalidad y el nombre de la relación.

Para definir el modelo el modelo en DSpace 7 se utiliza el concepto de «configurable entities». Esto se basa en no codificar las clases específicas del modelo (como por ejemplo *Persona*), sino que la implementación de un modelo de objetos específico es completamente configurable a través de un XML utilizado para determinar las relaciones entre las entidades. En este XML se pueden definir: los tipos de entidades que participan en la relación (una persona con una publicación, como se muestra en la figura 4, una persona con una organización, etc.), el *label* o nombre de la relación (isAuthorOfPublication, isPartOfOrganization), y la cardinalidad de la relación, con un mínimo y un máximo (por ejemplo, en el caso de la relación persona-publicación, el mínimo de publicaciones o máximo de publicaciones que puede tener una persona y el mínimo y máximo de autores que puede tener una publicación). Luego este XML se procesa y se vuelcan sus datos a la base de datos creando las entidades necesarias en el proceso (DSpace 7 - Configurable Entities, n.d.).

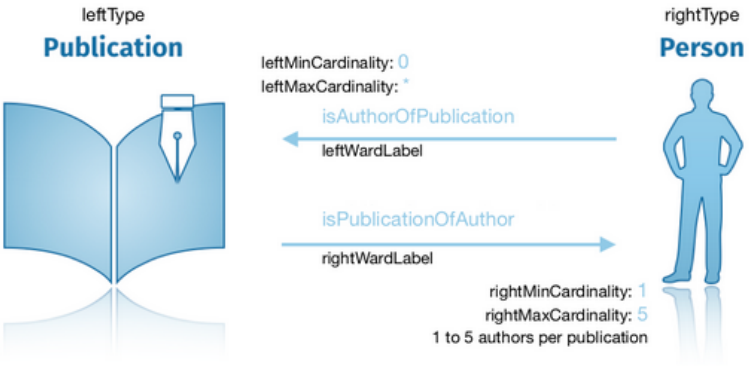


Figura 4. *Diagrama de relaciones entre una persona y una publicación*

```
                Table "entity_type"
 Column |          Type           | Modifiers
--------+-------------------------+-----------
 id     | integer                 | not null
 label  | character varying(32) | not null
```

Figura 5. *Tabla 'entity_type' que contiene el nombre de la entidad en DSpace 7*

```
                    Table "relationship_type"
       Column          |          Type          | Modifiers
-----------------------+------------------------+-----------
 id                    | integer                | not null
 left_type             | integer                | not null
 right_type            | integer                | not null
 left_label            | character varying(32)  | not null
 right_label           | character varying(32)  | not null
 left_min_cardinality  | integer                |
 left_max_cardinality  | integer                |
 right_min_cardinality | integer                |
 right_max_cardinality | integer                |
```

Figura 6. *Tabla 'relationship_type' que contiene los datos de una relación entre entidades en DSpace7*

```
                Table "relationship"
   Column    |  Type   | Modifiers
-------------+---------+-----------
 id          | integer | not null
 left_id     | uuid    | not null
 type_id     | integer | not null
 right_id    | uuid    | not null
 left_place  | integer |
 right_place | integer |
```

Figura 7. *Tabla 'relationship', la cual contiene los tipos de entidades que interactúan en la relación en DSpace 7*

CERIF

El modelo de datos que sustenta un CRIS se basa en un conjunto de entidades básicas definidas por el modelo del Formato Común Europeo de Información de Investigación (CERIF, por sus siglas en inglés) mantenido por la organización sin fines de lucro EuroCRIS.

El modelo de datos de CERIF permite una representación de las entidades de investigación, sus actividades, interconexiones y resultados, así como la flexibilidad suficiente para establecer relaciones semánticas. Este modelo consiste en proyectos, personas, organizaciones, publicaciones, patentes, productos, servicios y equipos, con relaciones temporales y basadas en roles. Las entidades involucradas en el modelo y sus relaciones se pueden apreciar gráficamente en la figura 8 (EuroCRIS | Main features of CERIF, n.d.).

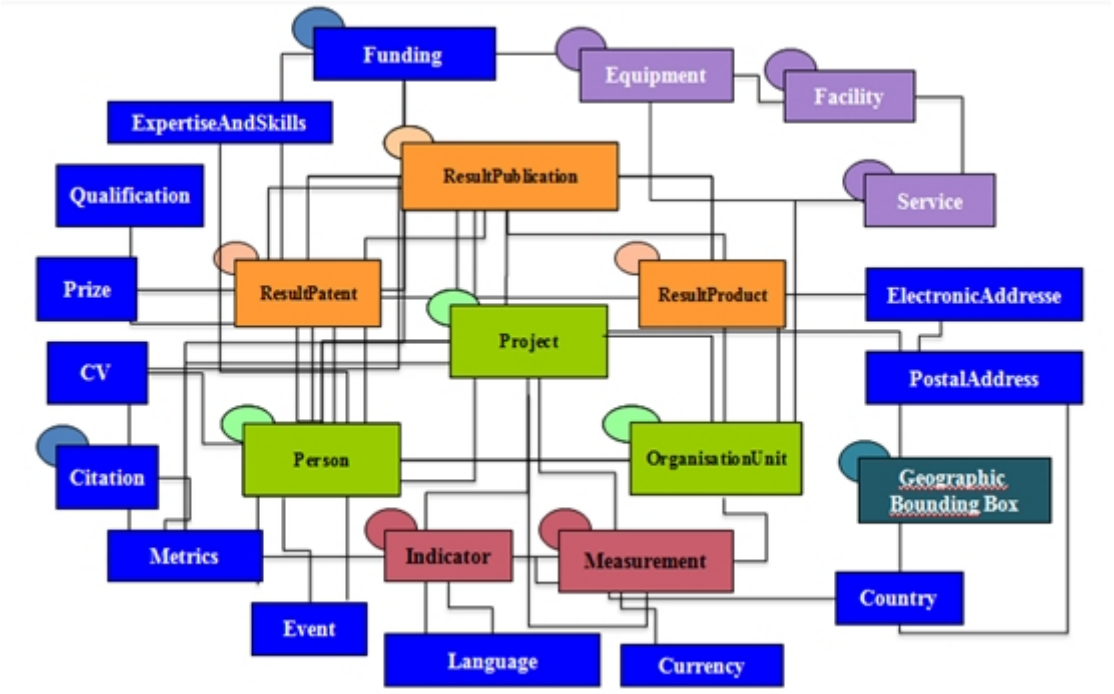


Figura 8. *Diagrama de relaciones entre las entidades del modelo CERIF*

Scopus

En Scopus, los perfiles de autores se generan automáticamente. El algoritmo agrupa los nombres de los autores bajo un identificador de autor de Scopus común, basado en un algoritmo que coincide con la filiación, la dirección, el área

temática, el título de la fuente, las fechas de citas de publicaciones e información de coautores (Krämer & Momeni, 2017).

En este caso el autor no se puede autenticar para tener un control total sobre sus datos pero puede solicitar modificaciones de sus datos personales que si hay errores, agregar o eliminar publicaciones, unificar perfiles duplicados y actualizar datos de filiación. Para solicitar los cambios se debe registrar en Scopus y seguir una serie de pasos para indicar cuáles son los cambios que se quieren realizar. Estas solicitudes se pueden realizar en el perfil propio o realizar en nombre de otras personas, y luego son analizadas por el equipo de Scopus para aprobarlas o no según corresponda.

En cuanto al modelo de datos, el análisis que se puede realizar corresponde al perfil público, ya que el código no se encuentra de forma abierta. Los datos que se incluyen son:

- *Nombre y apellido*
- *Filiación*
- *Documentos*
- *Identificador de ORCID*
- *Perfil de la red social académica Mendeley*
- *Área temática*

También se agregan algunos servicios, como la cantidad de citas, índice h, cantidad de coautores y suscripciones para alertar de las nuevas publicaciones del autor.

Capítulo 5 - Relevamiento de perfiles de autor en repositorios institucionales y sus servicios

Introducción

Para un análisis más en detalle de la implementación de los perfiles de autor en los repositorios institucionales y los servicios que se ofrecen alrededor de estos, se examinaron los primeros 200 repositorios del ranking web de transparencia de Webometrics. El relevamiento consistió en, primero, la búsqueda de aquellos repositorios, dentro de los 200 analizados, que tuvieran implementado el servicio de perfiles de autor para todos o algunos de sus autores, y luego, en una examinación minuciosa de todos los perfiles de autor encontrados, se identificó el software utilizado, los datos mostrados en cada una de las distintas implementaciones de los perfiles y los servicios que son provistos a partir de los datos recopilados en el perfil.

Este relevamiento permitió conocer el estado general del servicio de perfiles de autor en los repositorios institucionales; a partir de ello se puede obtener un estimado del porcentaje de repositorios que implementan estos perfiles y conocer el software y los servicios más utilizados dentro de los perfiles. Toda esa información es beneficiosa a la hora de decidir qué tipo de perfil de autor se quiere para una institución: al observar los distintos datos expuestos y servicios provistos se puede elegir con mayor facilidad y argumentos cuál puede ser el mejor modelo de datos a utilizar en un perfil para un repositorio determinado y los servicios a implementar en torno a él.

Repositorios con perfiles de autor

El reporte obtenido tras el relevamiento realizado muestra que de los 200 repositorios analizados sólo 28, un 14 %, implementa el servicio de perfiles de autor. Estos 28 repositorios son de diversas partes del mundo (de Argentina sólo se encuentra el repositorio digital de CONICET, CONICET Digital), y los hay tanto universitarios, como el Seoul National University's Institutional Repository S-Space o el Archive of Research de la Università degli Studi di Milano Institutional, como temáticos, por ejemplo, PhilPapers, repositorio de artículos y libros sobre filosofía. En la siguiente tabla se muestra el listado de los repositorios que implementan los perfiles de autor, con sus respectivos enlaces:

Repositorio	Url	País/Región
Europe PubMed Central	https://europepmc.org/	Global
Research Papers in Economics	https://ideas.repec.org/	Global
PhilPapers	https://philpapers.org	Global
Utrecht University Repository	https://www.narcis.nl/	Holanda
Università degli Studi di Milano Institutional Archive of Research	https://air.unimi.it/	Italia
National Chiao Tung University Institutional Repository	https://ir.nctu.edu.tw/	Taiwán
Digital CSIC	https://digital.csic.es/	España
RiuNet Repositorio Institucional Universidad Politécnica de Valencia	https://riunet.upv.es/	España
Seoul National University's Institutional Repository S-Space	http://s-space.snu.ac.kr	Corea del sur
Universitas Muhammadiyah Surakarta Digital Library	http://library.ums.ac.id/	Indonesia
Universitat Autònoma de Barcelona Dipòsit Digital de Documents	https://ddd.uab.cat/	España
Dépôt Institutionnel de l'Université Catholique de Louvain	https://uclouvain.be/	Bélgica
University of Ghent Institutional Archive	https://biblio.ugent.be/	Bélgica

Xi'an Institute of Optics and Precision Mechanics CAS Institutional Repository	http://english.opt.cas.cn/	China
Vietnam National University Digital Repository	http://repository.vnu.edu.vn/	Vietnam
Kazan Federal University Repository	https://repository.kpfu.ru/eng/	Rusia
Korea Advanced Institute of Science and Technology Open Access Self-Archiving System	https://koasas.kaist.ac.kr/	Corea del Sur
Queensland University of Technology Institutional Repository	https://eprints.qut.edu.au	Australia
Repositorio Institucional Universidad de Valladolid	http://uvadoc.uva.es/	España
Utah State University Digital Commons	https://digitalcommons.usu.edu/	EE.UU
Technical University of Denmark Online Research Database in Technology	https://orbit.dtu.dk/	Dinamarca
CONICET Digital	https://ri.conicet.gov.ar/	Argentina
University of Iowa Research Online	https://ir.uiowa.edu/	EE.UU
Archive Ouverte Université de Genève	https://archive-ouverte.unige.ch/	Suiza
University of Massachusetts Amherst Scholarworks	https://scholarworks.umass.edu/	EE.UU
Biblos-e Archivo Universidad Autónoma de Madrid	https://repositorio.uam.es/	España
University of Pennsylvania ScholarlyCommons	https://repository.upenn.edu/	EE.UU
Keio University Academic Resource Archive	http://koara.lib.keio.ac.jp/	Japón

Análisis de los perfiles de autor implementados

Del relevamiento se obtiene un análisis en detalle de los servicios implementados y las características más importantes en las 28 implementaciones del perfiles de autor del ranking de Webometrics. Se destacan las características más importantes de estos perfiles y se separan en 8 grupos: software utilizado, publicaciones, estadísticas, identificadores persistentes, datos personales, datos de filiación y redes sociales académicas; luego, se realizan reportes cuantitativos

de cada uno de estos grupos, la cantidad de repositorios que ofrecen esa característica o servicio, y particularidades de la implementación propia de cada repositorio. En el Anexo 1 se encuentra una tabla con el detalle de este análisis.

Software utilizado

Uno de los puntos relevados fue el del tipo de software que utilizan los repositorios. Dentro de este punto se diferenció el software utilizado para la implementación de los perfiles de autor con el utilizado para el repositorio en sí, que, como se detalla en el relevamiento, en algunas de estas implementaciones difieren.

Se encontró que DSpace es el software más utilizado entre estos repositorios, 10 de ellos lo usan, en detalle, uno de estos es en realidad un DSpace-CRIS, 3 utilizan DSpace en conjunto con otro software para el desarrollo de los perfiles de autor, y los restantes 5 realizaron personalizaciones de DSpace para la implementación de los perfiles. Después de DSpace, Digital Commons sigue en la lista de softwares usados con 4 repositorios, luego existen varios con solo una implementación, por ejemplo, EPrints, Drupal y algunos desarrollos *ad hoc*. Estos datos se pueden visualizar gráficamente en la figura 9.

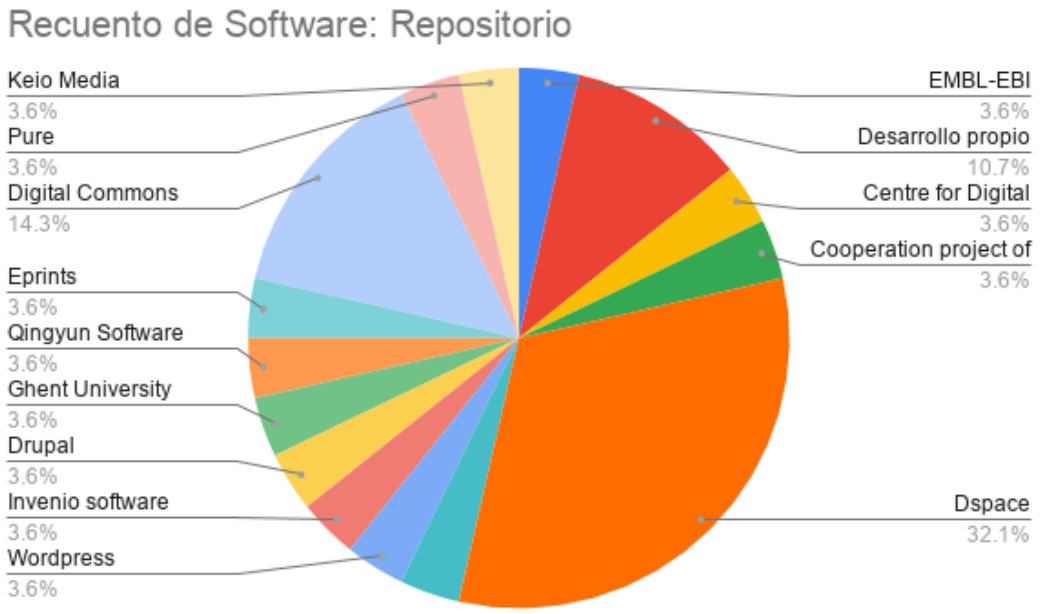


Figura 9. *Gráfico de los distintos softwares utilizados en los repositorios con perfiles de autor relevados*

Publicaciones

Agrupar todas las obras de un autor bajo una lista navegable es una de las características más comunes de los repositorios institucionales, y algo similar ocurre con los perfiles de autor: 26 de las 28 implementaciones de perfil de autor analizadas muestran las publicaciones del autor, generalmente en forma de listado con posibilidad de ordenarlo y realizar filtrados bajo diversos criterios (coautores, filiación o fecha de publicación, entre otros).

En contraste con el filtrado de publicaciones por autor que ofrecen la mayoría de los repositorios, el agrupamiento de las publicaciones de un autor en un perfil reduce las chances de incorporar entre sus elementos obras que no pertenezcan al autor a causa de ambigüedades en los nombres, puesto que puede ocurrir que varias personas aparezcan bajo nombres diferentes (sinónimos) o los distintos autores puedan tener nombres similares (homónimos) (Ferreira *et al.*, 2013). Los perfiles de autor mitigan este problema porque se cuenta con una mayor información sobre el autor y también con la posibilidad de definir variantes de su nombre, lo que permite a los administradores del repositorio, e incluso al mismo autor, detectar con mayor facilidad cuándo una publicación se encuentra introducida erróneamente en el listado de las obras de otro autor.

En muchos de los repositorios se observó también que se ofrecen servicios a partir del listado de las publicaciones, entre los que cabe destacar: la exportación de las publicaciones en distintos formatos (JSON, XML, CSV, PDF), la posibilidad de compartir o recomendar una publicación en redes sociales u otras plataformas, estadísticas sobre las publicaciones y la opción de exportar toda la bibliografía del autor, en conjunto con sus datos personales.

Estadísticas

Una de las mayores motivaciones de los miembros de las instituciones para depositar en Acceso Abierto es incrementar la visibilidad de sus publicaciones. En ese sentido, uno de los servicios de valor agregado que ofrecen varios de los repositorios es el reporte de estadísticas. Las estadísticas de uso suministradas

por los repositorios pueden ser muy interesantes y actuar como un fuerte incentivo para que los investigadores contribuyan con sus obras al acervo de los repositorios institucionales (COAR, 2013).

Se pueden visualizar estadísticas en 11 de los 28 repositorios relevados y estas van desde estadísticas de uso hasta estadísticas de las obras publicadas. Ejemplos de lo primero son la cantidad de visualizaciones o descargas en las obras de un autor, que son generalmente las estadísticas más mostradas; también esas mismas estadísticas pero agrupadas por año o región geográfica (país o ciudad) o la cantidad de visualizaciones a la propia página del perfil de autor. En cuanto a las estadísticas sobre los datos publicados se destacan el número de publicaciones del autor, agrupadas por tipo u organización, cantidad de estadísticas con los distintos coautores, redes de colaboración y número de citas al autor.

Por lo general, las estadísticas van acompañadas de gráficos de varios tipos (de torta, barra, temporales, etc.) que permiten una visualización más clara; en la figura 10 se puede ver un ejemplo de distintos gráficos estadísticos de un perfil de autor del repositorio de la Universidad Nacional de Vietnam.

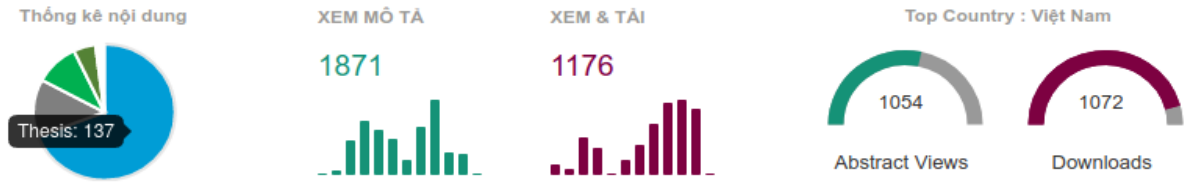


Figura 10. *Gráficos por tipo, cantidad de descargas y visualizaciones de las publicaciones de un autor en el repositorio de la Universidad Nacional de Vietnam*

Identificadores persistentes

Muchos de los repositorios analizados permiten la asociación de sus autores con algún identificador. Estos complementan los perfiles en la identificación unívoca de un autor, al ofrecer un código alfanumérico como identificador principal, lo que permite agrupar la producción e información del autor bajo este código, en vez de utilizar el nombre del autor, evitando así las confusiones y errores ocasionadas por la repetición de un mismo nombre entre autores distintos (Latif *et al.*, 2018).

De las 28 implementaciones de perfil de autor observadas, 9 permiten la integración con identificadores persistentes. Aunque esta integración no es exclusiva de los repositorios con perfiles de autor, del conjunto de los 172 repositorios restantes del relevamiento, que no poseen implementaciones de perfiles de autor, existen algunos que relacionan a sus autores con identificadores persistentes.

ORCID es el líder como proveedor de este servicio con la mayor presencia entre los repositorios con perfil de autor que integran identificadores persistentes, con 8 repositorios que permiten a sus autores asociar un ORCID a su perfil. Scopus ID lo sigue con 6 repositorios mientras que Researcher ID aparece integrado en 5 de ellos.

Datos personales

Detrás de las publicaciones y la filiación, la visualización de datos personales del autor es una de la características más comunes a todas las implementaciones de perfiles de autor. En este sentido, 22 de los 28 repositorios con perfil de autor muestran algún dato personal del autor (esto sin tener en cuenta el nombre de los autores que se muestra en la totalidad de los repositorios). Del conjunto de datos personales expuestos en los perfiles se destaca el e-mail y la foto, mostrados en 17 de los repositorios, lo siguen los datos profesionales, las áreas de interés y el teléfono del autor, mostrados en 10, 9 y 8 repositorios respectivamente; luego vienen otros datos como la biografía, la dirección y enlaces a páginas web o blogs del autor.

Datos de la filiación

La filiación de los autores es un dato con mucha presencia en los perfiles de autor: el relevamiento muestra que dentro de los 28 repositorios con perfil de autor estudiados, 24 exponen algún dato sobre la organización o la institución en la que el autor trabaja o de la que forma parte. En algunos casos, simplemente se muestra el nombre de la institución y en otros se detalla la dirección (a veces con la inclusión de mapas), teléfono y dependencias. Asimismo, existen repositorios

que optan por mostrar el historial de filiaciones y carrera profesional de sus autores, aunque la mayoría sólo da información sobre la filiación actual.

Redes sociales académicas

Las redes sociales científicas se han mostrado útiles como espacio colaborativo y de intercambio de conocimiento, y son relevantes para los autores a la hora de la construcción de su identidad digital, especialmente ResearchGate, Academia.edu y AutoresRedalyc. Estas redes científicas abiertas han evolucionado como sistemas que reconocen e interconectan perfiles (públicos o semipúblicos), comentarios, enlaces, áreas de interés, citas, reputación, popularidad y contenidos de todo tipo (Artigas & Casanova, 2020). Además, introducen nuevas métricas que pueden ser utilizadas en la evaluación científica (Froufe, 2016).

El relevamiento analizado tomó a las redes sociales académicas como un punto a estudiar en los perfiles de autor. Los resultados muestran 7 repositorios en sus perfiles de autor muestran enlaces a perfiles de los autores en alguna red social académica. Se pueden encontrar enlaces a Google Scholar, Researchgate, Publons, Dialnet, entre otras. La distribución exacta de los enlaces a las distintas redes sociales en estos 7 repositorios se pueden visualizar en la figura 11.

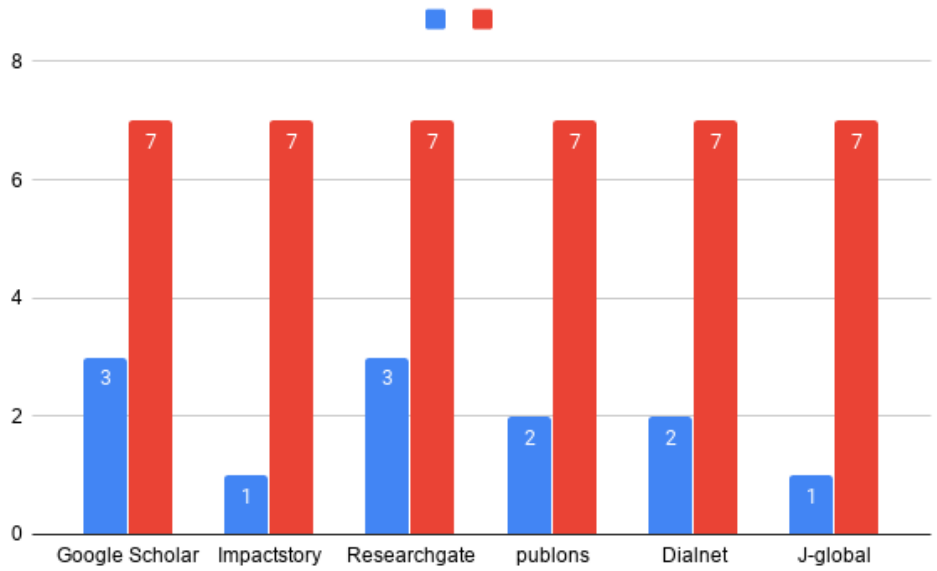


Figura 11. *Distribución de los enlaces a redes sociales en las distintas implementaciones de perfiles de autor encontradas en el relevamiento*

Conclusión

Como se mencionó antes, a pesar de que el servicio de perfiles de autor trae múltiples beneficios a los repositorios institucionales, desde mejoras en la visibilidad tanto del autor como del repositorio hasta la implementación de una serie de servicios de valor agregado en base a los datos de los perfiles, este servicio es poco implementado. Si bien el porcentaje de repositorios que implementan los perfiles de autor ha crecido a lo largo de los años, donde se observa que desde 2017 (Genovés, 2017) a 2020 la cantidad de repositorios del ranking web de Webometrics con este servicio aumentó de un 10 % a un 14 %, el aumento no es significativo, más si se tiene en cuenta el auge de los sistemas CRIS/RIMS. Sin embargo, con el lanzamiento de la versión 7 del software DSpace, que incorporará el servicio de perfiles de autor, se podría llegar a observar un incremento por parte de los repositorios en los años venideros, si es que los repositorios que utilizan DSpace (la mayoría de los repositorios institucionales) deciden actualizarse hacia la nueva versión.

En lo que respecta a los repositorios analizados que implementaron los perfiles de autor, el relevamiento permite ver que los servicios que se erigen alrededor de los perfiles son variados y difieren, dependiendo del repositorio analizado. Del conjunto de estos servicios se pueden destacar dos de ellos: el reporte de estadísticas y la integración con identificadores persistentes. Ambos servicios aportan un valor agregado al repositorio, en particular el reporte de estadísticas funciona como un incentivo para que los autores utilicen el repositorio y depositen allí sus trabajos, mientras que la integración con los identificadores persistentes permiten identificar unívocamente al autor e interoperar, tanto con el sistema proveedor del identificador como con otros sistemas que utilicen el mismo identificador para intercambiar y completar información de los autores, e incluso para encontrar obras de los autores de una institución que se encuentren dispersos por la web. Lamentablemente, los repositorios analizados no poseían registros o estadísticas que permitieran visualizar o sacar conclusiones sobre cómo la implementación de los perfiles de autor y los servicios en torno a ellos afectaron el uso del repositorio y la cantidad de obras depositadas por parte de los usuarios.

Capítulo 6 - Definición del modelo de datos, servicios y esquema de metadatos a implementar

Introducción

Una vez concluido el análisis de los modelos propuestos por los distintos sistemas que implementan algún servicio dedicado a perfiles de autores y el relevamiento sobre los repositorios mencionados en el capítulo anterior, se cuenta con suficientes herramientas para analizar ventajas y desventajas de cada servicio implementado, y así lograr una aproximación al modelo ideal que se desea implementar.

En un análisis rápido, se pueden destacar algunos aspectos en los servicios analizados: ya sea en el modelo propuesto por el software, o el sistema que implementa los perfiles de autores, hay una tendencia a realizar perfiles al estilo «curriculum vitae», con una buena presentación, datos personales y/o profesionales que permiten conocer las cualidades de un autor, su producción científica y hasta permiten comunicarse con él. También se ve que se suele conectar el perfil con los servicios más populares, como los proveedores de identificadores ORCID y SCOPUS ID o algunas redes sociales académicas como Google Scholar o ResearchGate. Por último, se puede observar que muchos de los sistemas analizados le dan una gran importancia a la visualización de estadísticas relacionadas con el investigador y su producción científica.

Dado este análisis se plantea diseñar un servicio de perfiles de autores en línea con las tendencias observadas, es decir, generar un perfil de autor con sus datos personales y profesionales, publicaciones, coautores, estadísticas, entre otras cosas que puedan enriquecer el perfil. Como ya fue mencionado, el contexto de esta tesina se da en el ámbito de los repositorios institucionales SEDICI y CIC

Digital, ambos implementados con el software DSpace en versiones 5 y 6 respectivamente, por lo que la implementación del modelo se decide hacer sobre DSpace. Se decide también utilizar la versión 7.0 de este software, debido a que ambos repositorios tienen programada una actualización a esa versión de DSpace, y, además, las versiones anteriores no tienen incorporada la funcionalidad de representar a los autores como entidades con su propia página a modo de perfil, como sí lo hace la versión 7. A su vez, por una cuestión de simplicidad, es importante partir de una base ya desarrollada por un equipo profesional, con una gran comunidad que lo sustenta y que se pueda adaptar fácilmente al software utilizado. También, como ya se dijo, el modelo propuesto por DSpace 7 se basa en el de Portland Common y de los sistemas CRIS, lo que permite tener diseños flexibles y extensibles de manera sencilla. De esta forma se pueden adaptar soluciones preexistentes y extender el modelo en base a los distintos requerimientos que surjan.

Definición del modelo de datos

A la hora de definir un modelo personalizado, DSpace 7, con su diseño de entidades configurables propone, a modo de buena práctica, realizar una serie de pasos a seguir para implementar el nuevo diseño de objetos:

1. Definir entidades del modelo
2. Identificar relaciones entre las entidades
3. Visualizar el modelo y definir un modelo de entidad-relación
4. Configurar relaciones

Definir entidades del modelo

A partir de la base del modelo de datos de DSpace 7 y de los distintos modelos relevados en el capítulo 4, se define que el modelo estará formado por entidades dinámicas unidas por una estructura de grafo y las entidades se conectarán entre sí a partir de relaciones. En base a los requerimientos planteados para diseñar el servicio, se pueden diferenciar tres objetos que interactúan en el modelo:

- *Publicaciones*: Esta entidad, como su nombre lo indica, hace referencia a toda la producción científica referida en el repositorio, ya sean artículos, tesis, libros o cualquier otro tipo de recurso soportado por el repositorio.

- *Personas*: Esta entidad se define para representar a los autores de las publicaciones del repositorio. Es importante destacar que, en el contexto de esta tesina, no todo autor se va a representar dentro del modelo, puesto que sólo se van a generar perfiles para los autores que cumplan con el requisito mínimo de pertenecer o haber pertenecido a la institución del repositorio: la UNLP en el caso de SEDICI y la CIC en caso de CIC Digital.
- *Organizaciones*: Esta entidad es la que va a representar a las instituciones.

Identificar relaciones entre las entidades

Una vez definidas las entidades se deben plantear las relaciones, lo que permitirá asociar a los autores con sus publicaciones e instituciones o a las publicaciones con las instituciones, entre otras conexiones. Estas conexiones potencian las opciones de navegación en el repositorio y permiten, por ejemplo, listar todas las publicaciones de un autor y acceder a cada una, o todas las instituciones a las que pertenece. Para este diseño se definen las siguientes relaciones:

- *isPublicationOfOrganization*: Define si una publicación pertenece a una institución. Esta relación permite listar las publicaciones de una organización o en el contexto de qué institución/es se realizó una publicación.
- *isAuthorOfPublication*, *isDirectorOfPublication* y *isEditorOfPublication*: Estas relaciones definen si la persona es autor/a, director/a o editor/a de una publicación y permite listar todas las publicaciones de un autor.
- *isOrgUnitOfPerson*: Define si una persona pertenece a una institución. Esta relación facilita el listado de instituciones para un autor dado, o todos los autores pertenecientes a una institución.
- *isCoauthorOfPerson*: Define si una persona tiene publicaciones con otra persona. Esta relación permite obtener los coautores de un autor dado.

Modelo de entidad-relación

El modelo de entidad-relación propuesto para el desarrollo de esta tesina puede verse gráficamente en la figura 12.

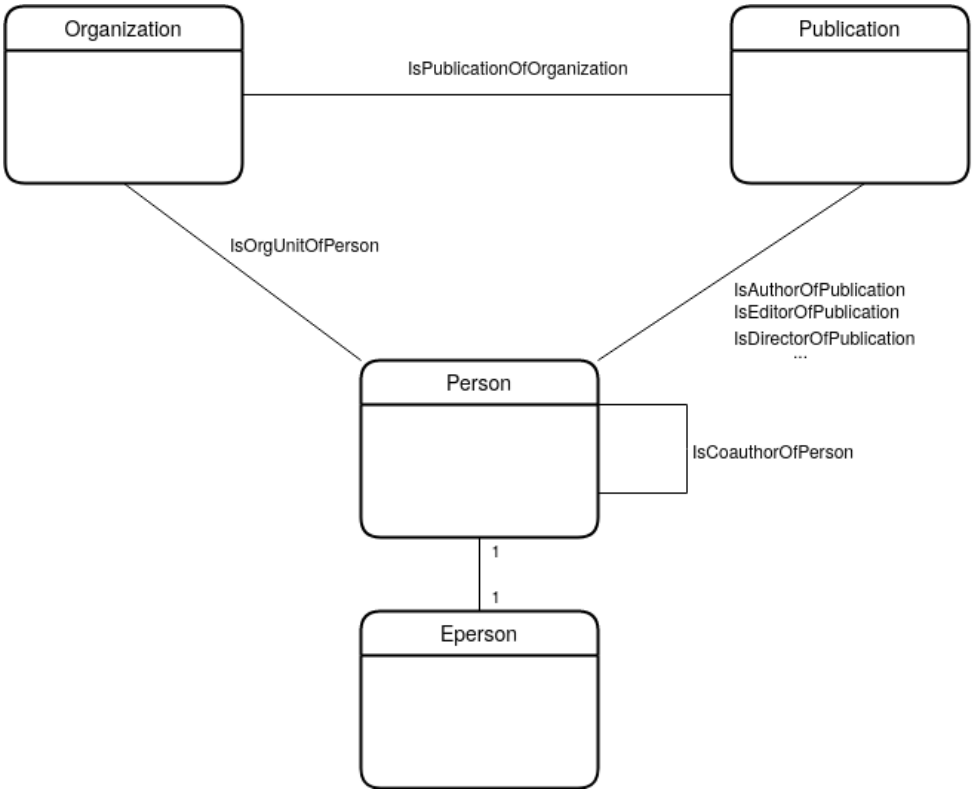


Figura 12. *Modelo de entidad-relación propuesto*

Una vez definido el modelo de entidad-relación es importante definir qué datos y servicios se van a incluir en el perfil de autor, para que en este contexto el análisis se base en la entidad Persona, que es la que representa el perfil en sí, mediante sus relaciones y atributos.

Del relevamiento de perfiles de autor analizado en capítulos anteriores se observa una tendencia a ver perfiles variados pero bastante completos, con datos personales, académicos, profesionales y algunos servicios. Considerando que una de las cosas que se busca es que el autor tenga una mayor relación e interacción con el repositorio, se desea implementar un perfil lo más completo posible, con servicios de interés para el investigador sin generar que la carga de los datos le demande mucho tiempo del autor y provoque un malestar con el servicio brindado.

Los atributos para implementar el perfil son:

- *E-mail*
- *Nombre completo*
- *Biografía*
- *Imagen*
- *Blog personal*
- *Identificadores persistentes*
- *Redes sociales académicas y personales*
- *Historial académico*
- *Áreas de investigación o de interés*

Tendrá relaciones con las siguientes entidades:

- *Organizaciones o instituciones en las que trabajó (filiaciones)*
- *Publicaciones realizadas*
- *Autores con los que realizó una publicación*

Configuración de las relaciones

Como se vio en el capítulo de modelos de datos (en la sección sobre DSpace 7), se utiliza el concepto de «Configurable Entities» para definir el modelo y sus relaciones. Para esto se define un XML con los tipos de entidades y su relaciones. En base al modelo planteado, se muestra en la figura 13 el archivo de configuración:

```
<relationships>

    <!-- Sample relationship types setup used for the entities development
     This file can be imported using the initialize-entities launcher -->
    <type>
        <leftType>Publication</leftType>
        <rightType>Person</rightType>
        <leftwardType>isAuthorOfPublication</leftwardType>
        <rightwardType>isPublicationOfAuthor</rightwardType>
        <leftCardinality>
            <min>0</min>
        </leftCardinality>
        <rightCardinality>
            <min>0</min>
        </rightCardinality>
        <copyToLeft>true</copyToLeft>
    </type>
    <type>
        <leftType>Publication</leftType>
        <rightType>Project</rightType>
        <leftwardType>isProjectOfPublication</leftwardType>
        <rightwardType>isPublicationOfProject</rightwardType>
        <leftCardinality>
            <min>0</min>
        </leftCardinality>
        <rightCardinality>
            <min>0</min>
        </rightCardinality>
        <copyToLeft>true</copyToLeft>
    </type>
    <type>
        <leftType>Person</leftType>
        <rightType>OrgUnit</rightType>
        <leftwardType>isOrgUnitOfPerson</leftwardType>
        <rightwardType>isPersonOfOrgUnit</rightwardType>
        <leftCardinality>
            <min>0</min>
        </leftCardinality>
        <rightCardinality>
            <min>0</min>
        </rightCardinality>
    </type>
    <type>
        <leftType>Publication</leftType>
        <rightType>OrgUnit</rightType>
        <leftwardType>isAuthorOfPublication</leftwardType>
        <rightwardType>isPublicationOfAuthor</rightwardType>
        <leftCardinality>
            <min>0</min>
        </leftCardinality>
        <rightCardinality>
            <min>0</min>
        </rightCardinality>
        <copyToLeft>true</copyToLeft>
    </type>
</relationships>
```

Figura 13. *Archivo de configuración de las relaciones*

Definición de los servicios

A la hora de definir los servicios que se implementarán en el armado de los perfiles se deben tener en cuenta algunos objetivos: por un lado, se intenta maximizar la visibilidad de la producción en el repositorio y, por otro, proveer un servicio de calidad, que sea útil para el investigador y le permita, por ejemplo, poder utilizar el perfil como una de sus identidades digitales. El relevamiento mencionado en el capítulo 5 es de gran ayuda para tratar de lograr estos objetivos, y si bien no están definidos de forma cuantitativa los resultados de la implementación de los servicios ofrecidos por los repositorios relevados, se pueden obtener varias ideas, estadísticas sobre los servicios más usados, y la forma en la que están implementados.

En base al relevamiento y los objetivos que se desean lograr, estos son los servicios que se consideran importantes para implementar en el perfil de autor:

Visualización pública del perfil de autor: incluye los datos personales y de investigador del autor, con sus publicaciones, filiaciones, etc. Esto permite tener los datos del autor en un mismo espacio, lo que ayuda a la hora de difundir la información completa del autor y refuerza las posibilidades de localizar al investigador a través de los motores de búsqueda, ya que en muchas ocasiones los usuarios desean contactarse con un autor y no encuentran el modo certero de hacerlo, por diversas circunstancias. En la figura 14 se presenta un ejemplo de

cómo se implementa la visualización del perfil de autor en RiuNet, el repositorio institucional de la Universidad Politécnica de Valencia.

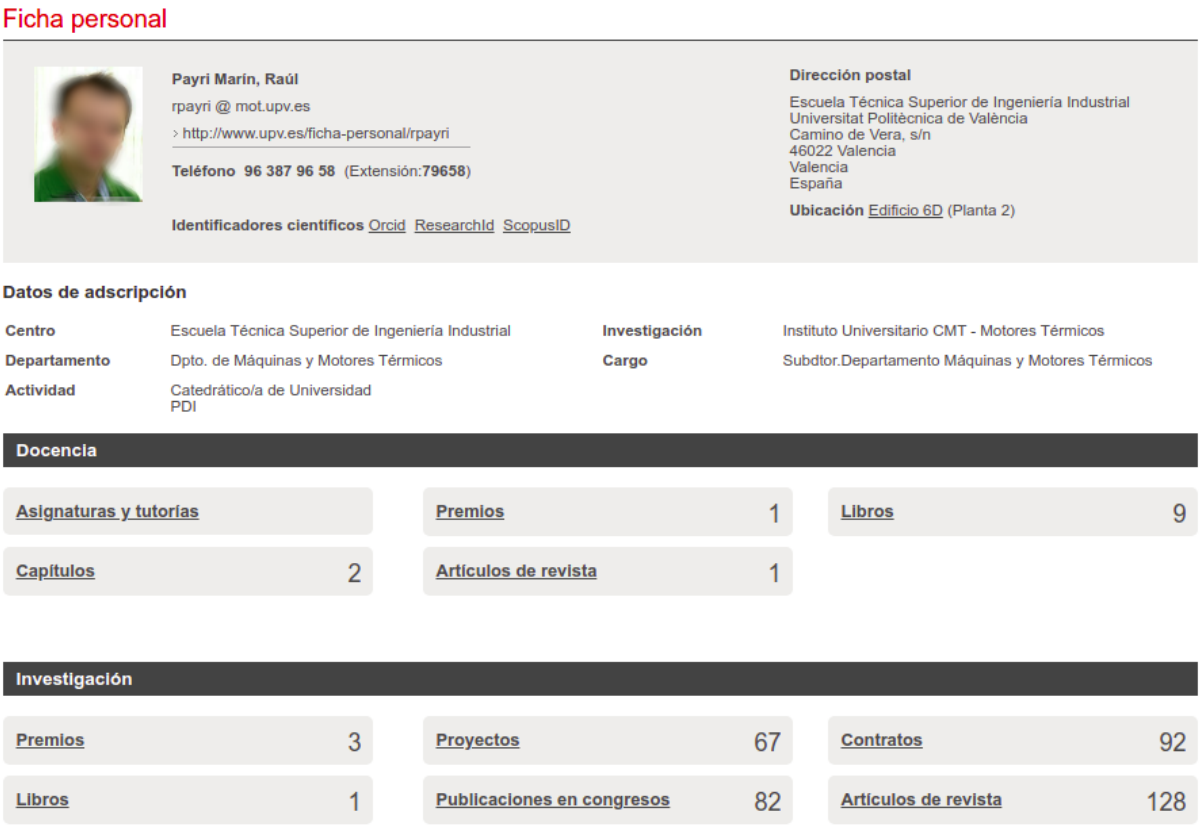


Figura 14. *Visualización del perfil de autor en RiuNet*

Navegación entre las relaciones de un autor, listarlas y poder realizar filtrados, y búsquedas en esos listados: Dentro de las relaciones podemos encontrar que un autor tiene coautores, publicaciones y filiaciones, lo que permite tener otra forma de llegar a la página de un autor al ir navegando por sus relaciones. En la figura 15 se puede ver cómo, en el repositorio del CONICET, CONICET Digital, desde un perfil de autor se puede acceder a una publicación, o a sus coautores:

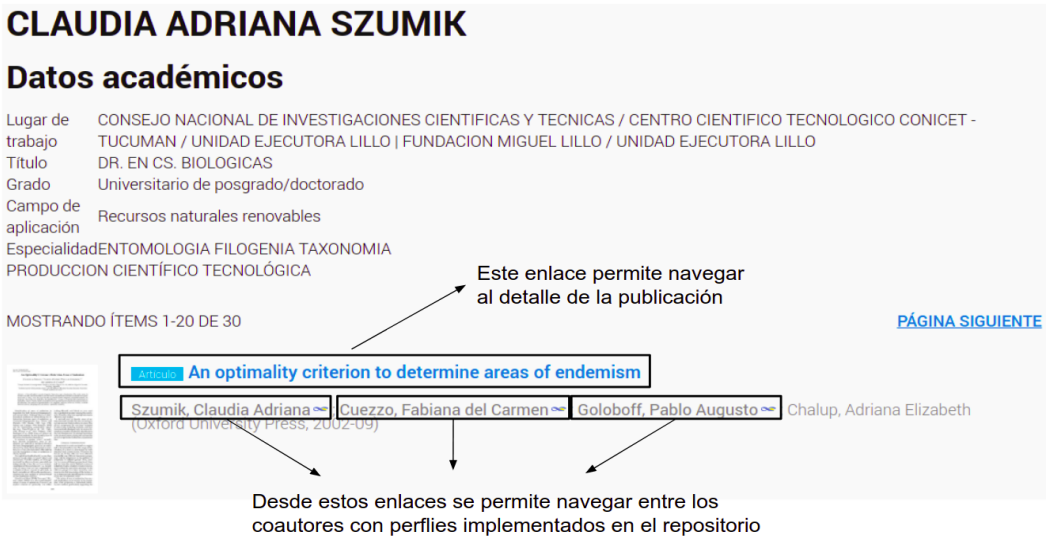


Figura 15. *Acceso a publicación a o sus coautores desde el perfil de autor en CONICET Digital*

Permitir a un autor autenticarse: DSpace 7 permite que un usuario cree una cuenta, inicie sesión y autoarchive sus publicaciones, pero no relaciona al usuario logueado con el autor o con su perfil. Por esto se plantea implementar un servicio que permita que un autor se pueda autenticar y relacionar el perfil con el usuario, como lo suelen hacer los proveedores de identificadores persistentes, o alguna de las redes sociales académicas. En la figura 16 pueden verse los ejemplos de ORCID y ResearchGate.

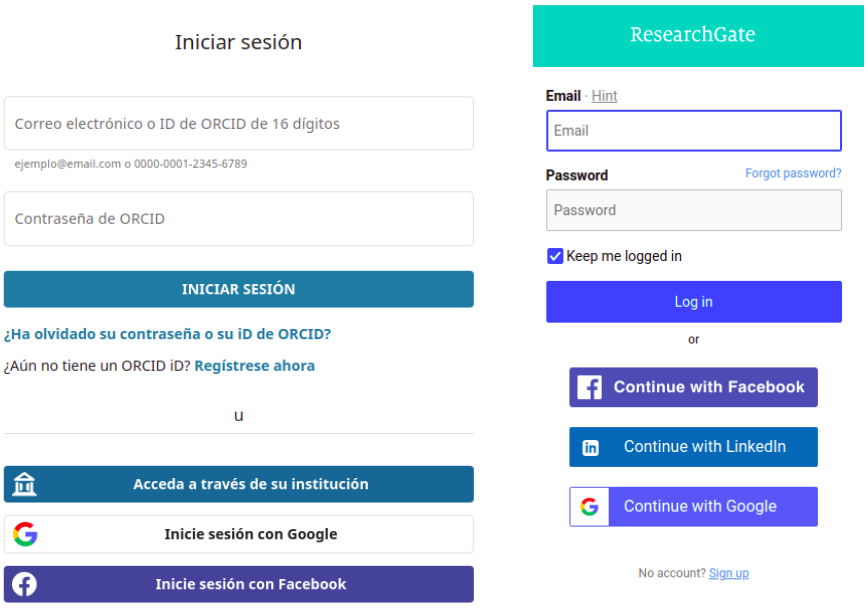


Figura 16. *Autenticación de usuario en ORCID y en ResearchGate*

Permitir a un autor administrar tanto sus datos de usuario como parte de su perfil de autor: Algunos datos del perfil no los podrá modificar directamente el autor sino que deberá solicitar a los administradores los cambios que considere necesarios. Dos ejemplos claros para ver son ORCID, que permite al autor modificar sus datos personales y configurar permisos para definir qué datos mostrar, y SCOPUS ID, en que el autor puede comunicarse con los administradores para avisar que algunos datos que no se pueden modificar son incorrectos desde un formulario. En la figura 17 se puede ver cómo ORCID implementa este servicio.

Figura 17. *Modificación de datos en el perfil de autor de ORCID*

Reporte o visualización de estadísticas en relación a la información y datos del autor, como sus publicaciones y coautores: esto agrega al repositorio un servicio de valor agregado que permite potenciar el perfil del investigador y tener un pantallazo general de la producción científica de cada autor, ofreciendo, por ejemplo, cantidad de publicaciones agrupadas por tipo de publicación, o bien cantidad de publicaciones a lo largo del tiempo o la cantidad de citas de una publicación. Además, este tipo de información es muy solicitada por los usuarios de los repositorios y los administradores no siempre pueden ofrecerla en tiempo y forma. En la figura 18, se puede ver cómo el Portal de Producción Científica de la Universidad Autónoma de Madrid (UAM) implementa su servicio de estadísticas:

Figura 18. *Estadísticas en perfil de autor del Portal de Producción Científica de la UAM*

Exportación del perfil de autor en diversos formatos como PDF, CSV y otros. En la figura 19 se puede ver cómo ORCID exporta a PDF los datos de un perfil:

Marisa De Giusti

https://orcid.org/0000-0003-2422-6322

Websites & Social Links

PrEBi - Proyecto de Enlace de Bibliotecas (http://prebi.unlp.edu.ar/staff/marisa-r-de-giusti/)

SEDICI - Repositorio de la Universidad Nacional de La Plata (http://sedici.unlp.edu.ar/discover?fq=author_filter%3Ade\+giusti%2C\+marisa\+raquel\|\|\|De\+Gi

CESGI - Centro de Servicios en Gestión de Información (http://cesgi.cic.gba.gov.ar)

Country

Argentina

Keywords

Digital Libraries, Institutional Repositories, Library Science

Other IDs

Scopus Author ID: 6601944550 (http://www.scopus.com/inward/authorDetails.url?authorID=6601944550&partnerID=MN8TOARS)

ResearcherID: K-2941-2015 (http://www.researcherid.com/rid/K-2941-2015)

Employment (5)

Comisión de Investigaciones Científicas: La Plata, Buenos Aires, AR

1995-10-01 to present | Main Researcher

Employment

Source: Marisa De Giusti

Universidad Nacional De La Plata Facultad de Ingenieria: La Plata, Buenos Aires, AR

1994-03-01 to present | Professor

Employment

Source: Marisa De Giusti

Universidad Nacional de la Plata: La Plata, Buenos Aires, AR

1994-03-01 to present | Professor (Computer Science)

Employment

Figura 19. *Exportación del perfil de autor a PDF en ORCID*

Exponer el perfil en algún formato para interoperar, como OAI, OpenSearch y otros. Los repositorios no deben permanecer aislados, dado que su valor real reside en su potencial para conectarse entre sí con el fin de construir una red de repositorios que facilite el acceso unificado a la gran cantidad de investigación abierta y a los materiales relacionados con ella, de tal manera que se abran nuevas formas de trabajar con la información (Fernández & Ferreras, 2013). Asimismo, los repositorios pueden generar mecanismos de interoperabilidad con otros tipos de sistemas informáticos vinculados a la gestión y generación del conocimiento, por ejemplo el sitio web de una revista científica o el campus virtual para educación a distancia de una universidad (De Giusti et al., 2013). Por este motivo es importante evaluar y generar distintos formatos de interoperabilidad que permitan aumentar la visibilidad de la producción o exponerla en otros sistemas, como en el proyecto de visibilidad web de la Universidad Nacional de La Plata, donde se utiliza el protocolo OpenSearch para recuperar la producción científica de las distintas unidades académicas de la institución y se exponen en sus sitios web (Villarreal, 2017). En la figura 20 se puede ver cómo se recupera, a través del protocolo OpenSearch, la producción de un autor desde el repositorio SEDICI y se muestra en un sitio web de un centro o instituto de la UNLP.

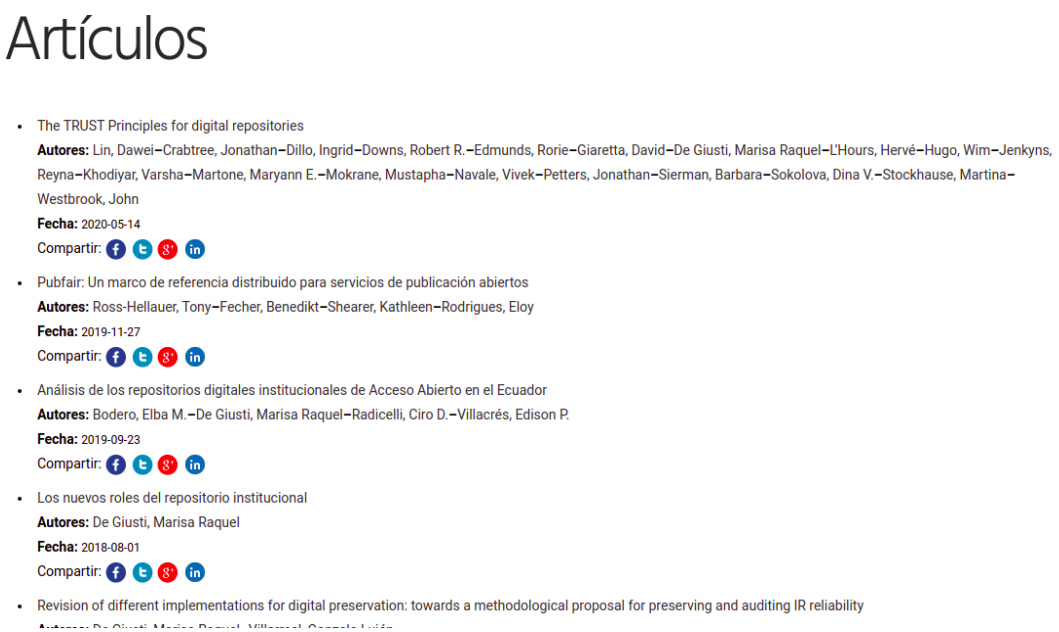


Figura 20. *Producción científica de un autor recuperada desde el repositorio SEDICI hacia un sitio web mediante protocolo OpenSearch*

Código QR con URL al perfil. En la figura 21 se puede ver cómo ORCID permite obtener un código QR para un perfil de usuario.

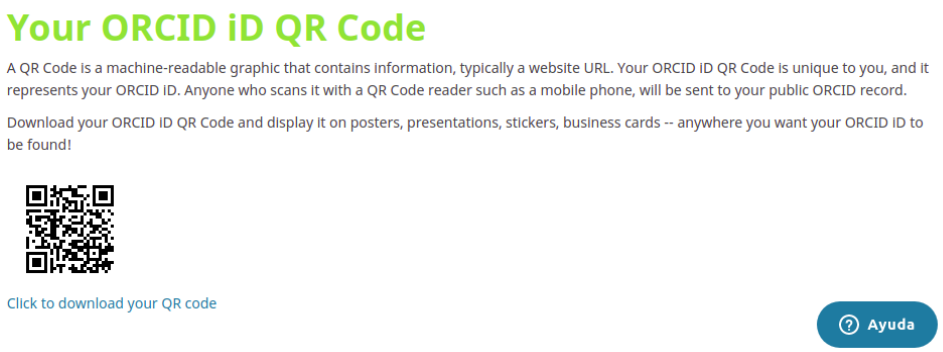


Figura 21. *Obtención de código QR para un perfil de autor en ORCID*

Búsqueda de autores a partir de algún identificador persistente en el repositorio. En la figura 22 se puede ver cómo el repositorio Europe PubMed Central permite realizar la búsqueda de autores desde algún identificador persistente, como ORCID:

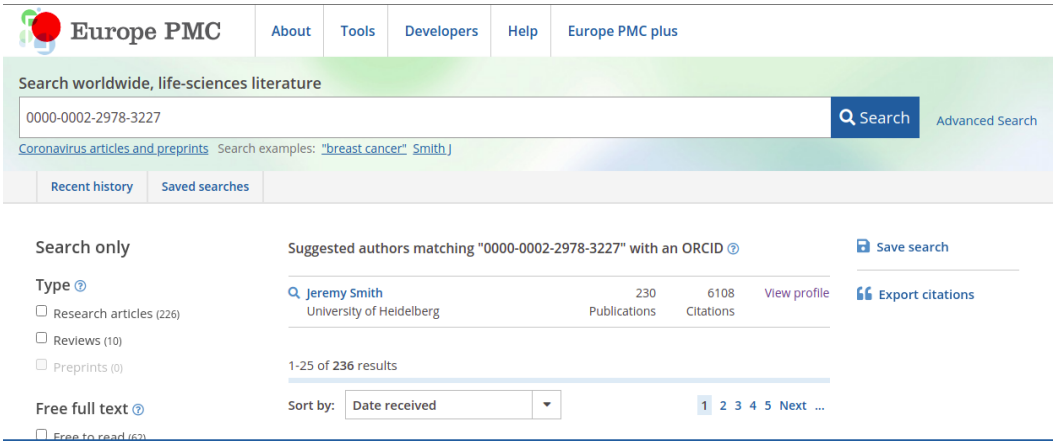


Figura 22. *Búsqueda de autores a partir de identificadores persistentes (ORCID) en Europe PMC*

Esquema de metadatos

Los metadatos se emplean para describir el contenido y otras características de los recursos digitales, posibilitando a una persona o máquina la búsqueda, gestión o recuperación de los datos. Los metadatos son datos secundarios como el autor,

el título, las palabras clave, el resumen, la fecha, u otros, que describen los datos primarios o recursos de información (García, 2010). Es decir, los metadatos permiten añadir información a un recurso digital con el fin de facilitar su gestión, recuperación, interoperabilidad o entendimiento del registro. Los conjuntos de metadatos relacionados con un recurso tienen una estructura bien definida y un conjunto de reglas para su uso, denominada esquema de metadatos (Vargas Arcila, 2016). Existe una gran cantidad de esquemas de metadatos genéricos definidos para distintos tipos de recursos que ayudan a normalizar la representación de un objeto digital, lo que facilita la tarea a la hora de interoperar con otros sistemas o de comprender un recurso. Un ejemplo es el modelo de Dublin Core, que propone un conjunto de elementos básicos, como título, autor, tipo de recurso, palabras clave, entre otros que permiten representar un objeto digital.

El modelo de datos propuesto en las secciones anteriores propone dos tipos de objetos digitales a representar dentro del repositorio, sin tener en cuenta las publicaciones, que es un objeto digital que ya se encuentra representado dentro de los repositorios; por un lado, se encuentran las organizaciones, que representan las filiaciones a las que pertenecen los investigadores y por el otro se tiene a los autores, que son el objeto digital principal a implementar en esta tesina. DSpace 7 implementa estas dos entidades y extiende los esquemas *Organization* y *Person* de Schema.org, una comunidad abierta, fundada por Google, Microsoft y otras grandes empresas con el fin de crear, promover y facilitar el uso de esquemas de metadatos en internet. Los esquemas propuestos por Schema.org abarcan gran parte del modelo presentado en esta tesina, por lo que se decide utilizar y extender los ya nombrados. A continuación se ofrece el detalle del esquema de metadatos a implementar.

Nombre	Metadato	Descripción
Dirección	*person.address*	Domicilio de una persona
Biografía	*person.biography*	Biografía o breve resumen de vida de la persona

Fecha de nacimiento	*person.birthDate*	Fecha de nacimiento de la persona
DNI	*person.dni*	Documento de identidad de la persona
E-mail	*person.email*	E-mail de la persona
Apellido	*person.familyName*	Apellido de la persona
Nombre	*person.givenName*	Nombre de la persona
URL de Google Scholar	*person.identifier.gsid*	Dirección web del perfil de Google Scholar de la persona
ORCID	*person.identifier.orcid*	ORCID del autor
Researcher ID	*person.identifier.rid*	Researcher ID del autor
Scopus ID	*person.identifier.scopus-author-id*	Scopus ID del autor
Título, profesión o trabajo	*person.jobTitle*	Título profesional y/o profesión/trabajo
Palabras clave	*person.keyword*	Palabras clave o áreas de interés del autor
Idiomas	*person.knowsLanguage*	Idiomas que domina el autor
ORCID URL	*person.orcid.url*	URL al perfil de ORCID del autor
Publons URL	*person.rid.url*	URL al perfil de Publons del autor
Scopus URL	*person.identifier.scopus-author-id.url*	URL al perfil de Scopus del autor

Teléfono	*person.telephone*	Teléfono de la persona
Páginas web	*person.webpage*	URL a páginas web y perfiles de redes sociales del autor
Variantes de nombre	*person.nameVariant*	Variantes del nombre del autor

Capítulo 7 - Recopilación, análisis y normalización de los datos de autores

Introducción

Uno de los puntos más importantes en la construcción de perfiles de autor en un repositorio institucional son los datos de los autores: es fundamental contar con información correcta y validada de los autores a la hora de implementar los perfiles. En el ámbito de esta tesina, estos datos se corresponden con los de investigadores pertenecientes a la Universidad Nacional de La Plata. SEDICI, el repositorio institucional de la UNLP, contiene en su base de autoridades datos de miles de autores y entre estos datos se encuentran su nombre, apellido, e-mail, categoría de investigador y la facultad o el departamento, centro y/o instituto dentro de la universidad al que pertenecen. Sin embargo, para muchos de estos autores esta información está incompleta, desactualizada o incluso puede que no esté correctamente validada. Si se quiere generar perfiles de autores contando sólo con esa información, éstos pueden llegar a exponer información errónea y no contarán con información valiosa de los autores como algún identificador persistente, sus redes sociales académicas, su cargo dentro de su departamento, las áreas de investigación o su título profesional. Por este motivo es importante generar y gestionar una base de datos de autores controlada y normalizada en el repositorio, que permita tener información correcta y de valor agregado, ofrecer distintos servicios para los investigadores y hasta interoperar con otros sistemas que lo requieran.

Toda esa información sobre los autores puede ser completada desde distintas bases de datos de sistemas orientados a la investigación académica en donde los autores de la universidad tengan un perfil de autor o dispongan allí de información actualizada o complementaria a la información que se tiene en SEDICI. Además, estas bases de datos podrían contener incluso datos de autores que no existen en la base de datos de SEDICI, lo cual permitiría generar incluso una mayor

cantidad de perfiles e incrementar la cantidad de documentos en el repositorio. Entre las potenciales bases de datos de las que se puede extraer esta información se encuentras los proveedores de identificadores persistentes (Scopus, ORCID, Publons), las redes sociales académicas (ResearchGate, Academia.edu, Google Scholar), bases de datos institucionales de la UNLP (el sistema de recursos humanos SIU Mapuche, la base de participantes de los proyectos de extensión de la UNLP) y otras bases de datos de instituciones con autores relacionados a la universidad (la base de datos de autoridades de CIC Digital o los autores con filiación UNLP del repositorio del CONICET).

Esta recopilación e integración de los datos de autores UNLP desde distintas fuentes se puede dividir en una serie de pasos:

1. Como primer paso, generar una lista de las bases de datos que posean información de autores de la universidad.
2. A partir de estas bases de datos, si es posible, obtener la mayor cantidad de información posible de los autores de la UNLP mediante distintas técnicas de recuperación de datos, como puede ser algún API REST, *web scraping* o simplemente solicitando a la institución pertinente una copia de la base de datos obtener los datos.
3. Una vez recopilada toda la información, se debe descartar aquella que es superflua para la creación de los perfiles de autor, a saber, cantidad de citas de un autor en las publicaciones de esa base de datos, ranking de los autores en cada una de las bases de datos, ID utilizados sólo en forma interna, etc.
4. Una vez obtenidos los datos de interés de cada base se deben limpiar y normalizar ciertos valores para unificar criterios y facilitar el posterior proceso de los datos: por ejemplo, a la hora de analizar el nombre y apellido de los autores, se pueden encontrar varias formas de representarlos, en dos columnas, primero el apellido, primero el nombre, etc.
5. Se debe realizar una unificación y detección de duplicados de los datos traídos desde las distintas bases de datos. Estos dos últimos procesos sirven para detectar un mismo autor en las distintas bases de datos y así

poder integrar la información recopilada desde todas ellas en un único registro. En este paso, al momento de encontrar un duplicado, se deben tomar algunas decisiones: dados dos datos distintos de una misma categoría para el mismo autor, cuál tiene mayor prioridad o si ambos deben permanecer en el registro. Por ejemplo, si para el autor «Juan Pérez», desde Scopus se recopiló el e-mail «juanperez@example.com», pero desde ResearchGate se obtuvo el mail «jperez@example2.com» se tiene que decidir cuál de estos dos e-mails prevalece o si se desea conservar ambos.

6. Para finalizar, se deben discriminar aquellos autores a los que finalmente se les desea crear un perfil, dado que no es deseable crear un perfil para todos los autores recopilados, pues algunas de las personas del documento podrían no tener publicaciones académicas o quizás no se tiene información suficiente de un autor como para crearle un perfil, por lo que se debe, bajo algún criterio, seleccionar los autores adecuados para la generación de un perfil.

Recopilación de datos de autores UNLP desde distintas fuentes

Al momento de analizar las distintas fuentes para obtener información de los autores, se observa que no existe una única base de datos que contenga información normalizada y completa de todos los autores cuya investigación, lugar de trabajo, o dependencia institucional sea la Universidad Nacional de La Plata y que, además, esa base de datos sea de acceso abierto. En consecuencia, si se quiere obtener la mayor cantidad posible de datos de autores UNLP estos se tienen que recopilar desde distintas fuentes. Para ello se realizó un relevamiento de las principales bases de datos de autores a nivel institucional, nacional y mundial, en el que se destaca a grandes rasgos qué se puede obtener de cada fuente y de qué forma se pueden obtener los datos, entre otras cosas.

En la dirección PREBI-SEDICI de la UNLP funcionan varios proyectos que por su naturaleza se encuentran en continuo contacto con muchas de las fuentes ya mencionadas para obtener, normalizar y cruzar los datos de los autores. Por este

motivo se generó un documento colaborativo con los equipos de dichos proyectos para que puedan aportar distintas fuentes de datos, instrucciones sobre cómo extraer la información, contactos para solicitar las bases de datos institucionales, o hasta, en algunos casos, bases ya procesadas, como en el caso de Google Scholar, en donde existe un procesamiento ya realizado por el equipo encargado de recuperar documentos para integrar al repositorio, el cual contaba con un archivo .csv con aproximadamente 5000 autores pertenecientes a la Universidad Nacional de La Plata.

En el Anexo 2 se encuentra la tabla con el listado de las bases de datos desde donde se extrajo información, junto con la cantidad de presuntos autores UNLP que se encontraron en cada una de ellas y el método utilizado para recopilar los datos, si se decidió incluir o no los datos obtenidos en la base de datos final y algunos comentarios al respecto. En la siguiente tabla se observa un resumen de las bases de datos que se incluyeron en la base de datos final:

Descripción de la fuente	URL	Autores UNLP encontrados
Base de autoridades SEDICI	http://sedici.unlp.edu.ar/	24276
Google Scholar	https://scholar.google.com/	4496
Participantes en Proyectos de extensión	http://proyectos-extension.unlp.edu.ar/	14038
Integrantes Proyectos SECYT 2013		3791
ResearchGate	https://www.researchgate.net/institution/Universidad_Nacional_de_La_Plata/members	3340
Scopus	https://www.scopus.com/freelookup/form/author.uri?zone=TopNavBar&origin=NO%20ORIGIN%20DEFINED	3399
ORCID	https://orcid.org/orcid-search/search?institution=%22national%20university%20of%20la%20plata%22	2562
Publons (Researcher ID)	https://publons.com/ https://www.researcherid.com/	137

RePEc Authors	https://ideas.repec.org/i/e.html	52

Entrecruzamiento y normalización de los datos recopilados

Una vez que se obtuvieron los datos desde todas las fuentes deseadas, el siguiente paso consistió en normalizar y limpiar cada una de las fuentes para unificar la información y tener como resultado una única base de datos en un único formato, en vez de en distintos archivos con formatos que varíen uno de otro. De las fuentes obtenidas durante este proceso, se puede destacar que los formatos de los archivos, en general, son .csv, .json y hasta en algún caso se obtuvo alguna fuente en .xml. Se pueden encontrar múltiples herramientas y librerías que facilitan la manipulación de datos en estos formatos, ya sea para unificar, eliminar o filtrar registros de las fuentes. En cuanto a la elección del formato la base de datos final (base de datos normalizada y unificada de todas las fuentes), se decidió utilizar el formato .csv, que no solo permite una gran facilidad a la hora de manipular los datos, sino que también es compatible con la mayoría de las herramientas a utilizar durante este proceso.

Con el uso de la herramienta OpenRefine, que permite la limpieza y transformación de datos desde distintos formatos, se unificaron todos los datos de las distintas fuentes en un único archivo .csv. Para esto, primero se convirtieron todos los archivos en formato JSON o XML a CSV, proceso simple de hacer con la herramienta citada, que al abrir uno de estos archivos permite visualizarlo en forma de tabla y luego exportarlo en el formato deseado.

Una vez con todos los archivos en formato .csv, se prosiguió a normalizar los nombres de las columnas o campos. Se siguió el mismo criterio de nombres para todos los archivos; por ejemplo, si para el campo o columna donde se encuentran todos los e-mails de los autores, en el archivo recopilado desde Scopus esta columna se llamaba «Scopus email» pero en el de SEDICI se llama «author email», se decidió renombrarlos solo por «email». Con la misma lógica se renombraron todas las columnas que contenían el mismo tipo de dato pero que poseían distinto nombre en los diferentes archivos y entre estas columnas repetidas se

encontraban el nombre y apellido del autor, la filiación, ORCID, Scopus ID, e-mail, entre otros.

Además de normalizar los nombres de las columnas, también se realizó una normalización de su contenido, para que en todos los archivos los datos de un mismo tipo siguieran un único formato. Uno de los casos en el que se aplicó esta normalización fue en el de las celdas multivaluadas, es decir, celdas en las que hay varios valores de un mismo dato de un autor. Los distintos valores dentro de este tipo de celdas podían estar separados por diferentes caracteres dependiendo el archivo; por ejemplo, en el archivo proveniente de Scopus si un autor tiene varios email se lo podría haber separado con el carácter (|) de la siguiente manera: «email1|email2», mientras que en el archivo proveniente de Google Scholar se lo podría haber separado con el carácter (#). En este tipo de situaciones se decidió separar los valores por un mismo carácter en todos los archivos. Otro ejemplo que se da a la hora de la normalización de contenido tiene que ver con el nombre y apellido de los autores, en las distintas fuentes se pueden encontrar muchas variantes en la forma de representarlos; en algunos casos en columnas separadas, en otros casos el valor se formaba con el apellido primero, seguido por una coma (,) y luego por el nombre, o se comenzaba con el nombre. También para estos casos se decidió unificar un criterio para normalizar el contenido de todas las fuentes.

También se eliminaron columnas de los distintos CSV que no aportaban información valiosa para la unificación de datos final. Estas columnas, por lo general, contenían datos del autor pero relacionados de forma muy estrecha con la fuente desde donde se obtuvieron, como el ID interno de un autor en el repositorio RePEc de publicaciones económicas, la cantidad de publicaciones que un autor posee en ORCID, o la posición de un autor en el ranking de autores de Scopus.

Finalmente, luego de normalizar el nombre de las columnas, se unió el contenido de los archivos en un único proyecto con OpenRefine. Si con esta herramienta se abren varios archivos CSV a la vez, automáticamente se genera un proyecto que une todos los archivos en una única tabla, con tantas columnas como distintas columnas hay entre los archivos, y con tantas filas como la suma de las filas de todos los archivos abiertos, y rellena con blanco las celdas nuevas creadas

en el proceso, además de agregar una nueva columna que indica de qué archivo proviene cada fila. Es decir, si se tiene un CSV A con la siguiente estructura:

Nombre	Apellido	Email	Archivo origen
Juan	Perez	jp@unmail.com	archivo_a.csv

Y un CSV B como el siguiente:

Apellido	Institución	Archivo origen
Gonzales	UNLP	archivo_b.csv

Entonces, al abrirlos ambos con OpenRefine, se genera la siguiente tabla:

Nombre	Apellido	Email	Institución	Archivo origen
Juan	Perez	jp@unmail.com		archivo_a.csv
	Gonzales		UNLP	archivo_b.csv

Esta tabla resultante es posible exportarla en una variedad de formatos; en el caso analizado, fue CSV. El resultado de esta unificación de todas las fuentes recopiladas fue entonces un archivo CSV, como se dijo, con 61 columnas (incluyendo la columna que agrega OpenRefine de manera automática que indica el archivo de origen de cada fila) y 57504 filas, resultado de la unión de los 2562 autores extraídos desde ORCID, de los 3399 desde Scopus, 137 de Publons, 4496 de Google Scholar, 24276 desde la base de autoridades de SEDICI, 3791 desde la base de datos de integrantes de proyectos SECYT UNLP, 1413 y 14038 mediante las bases de datos de los participantes en proyectos de extensión UNLP, 52 desde el repositorio RePEc y 3340 desde ResearchGate (ver figura 23). Algunas de las columnas del CSV unificado quedaron con datos exclusivos de la fuente de la que fueron extraídos (por ejemplo, ID del autor en Google Scholar, el CUIT del autor

obtenido solo desde SEDICI), y otras con datos compartidos entre algunos de los archivos (como el nombre de los autores, su e-mail, ORCID, etc.). Sin embargo, si un mismo autor aparecía repetido en varias de las fuentes unificadas, entonces su información completa quedaba esparcida, y parte de su información duplicada, en varias filas dentro del archivo resultante; una fila puede contener su e-mail y filiación extraídos de ORCID mientras que en otra su cantidad de citas en Google Scholar o su título profesional, recopilados desde la base de autoridades de SEDICI y ambas filas pueden poseer su nombre completo y fecha de nacimiento.

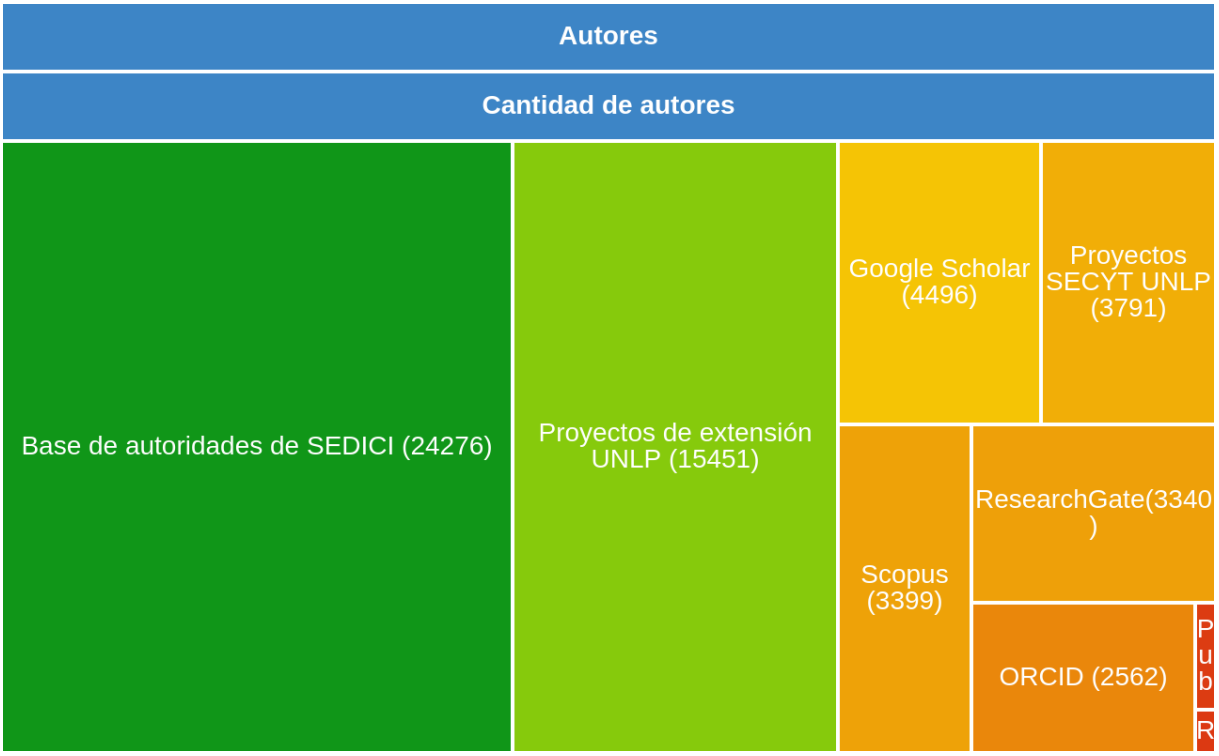


Figura 23. Gráfico con la cantidad de autores recolectados por fuente.

Deduplicación de los datos

Luego de haber generado el archivo CSV con todos los datos obtenidos desde las distintas fuentes unificadas (el cual se nombrará como «CSV unificado»), la información de muchos de los autores quedó distribuida, o en unos cuantos casos duplicada, en varias filas. Esto ocurre a causa de que se puede haber extraído información de un mismo autor desde distintas fuentes, y la información del autor proveniente de cada una de esas fuentes quedó en una fila distinta del archivo.

Por lo tanto, no se pueden generar los perfiles a partir de los datos crudos del CSV unificado, pues eso daría lugar a que se generen varios perfiles para un mismo autor; entonces, se necesita realizar un proceso de deduplicación de los autores, agrupando toda la información de un mismo autor en una misma fila y eliminando aquellos datos o filas que se repitan.

Este proceso de deduplicación es complejo y requiere tener en cuenta factores como los nombres homónimos o la poca información que se pudo extraer sobre los autores desde algunas fuentes y que no permiten identificar unívocamente a un autor. En muchos casos, solo se tiene el nombre del autor y la filiación, y es común que la filiación sea solo «Universidad Nacional de La Plata» (filiación que los autores tienen en común ya que se buscaron todos los autores de esta institución); en casos así, si se encuentra más de un autor obtenido desde fuentes distintas con el mismo nombre y la misma filiación (o una filiación que pertenezca a la UNLP) se dificulta mucho saber, si no se cuenta con datos adicionales de los autores, si ambos son realmente la misma persona o si son personas distintas pero homónimas que pertenecen a la UNLP. El caso mencionado, de todas maneras, tiene poca relevancia, porque al deduplicar un autor solo con el nombre y filiación, no suma datos nuevos de una tupla a la otra, por lo cual no va a sumar datos erróneos (a fines prácticos, son el mismo autor). El problema surge cuando las filas que se deduplican vienen de fuentes con más datos, por ejemplo si la primera tupla pertenece a un autor que además tiene como dato el ORCID o el DNI (datos que representan unívocamente a un autor), y la segunda también tiene datos adicionales, como Scopus ID o algún otro, se debería asegurar que las dos tuplas hagan referencia al mismo autor, y no generar una tupla con datos erróneos, pertenecientes a más de una persona.

A partir de este análisis fue que se decidió comenzar a deduplicar las filas del CSV a partir de los identificadores unívocos de un autor (ORCID, e-mail, Scopus ID, DNI, etc.), los cuales permiten (a menos que la fuente desde donde se extrajeron los datos no sea confiable) marcar como duplicados, con una gran exactitud, a los autores que tienen un mismo identificador; una vez realizado este paso, se debe realizar la deduplicación por nombre y filiación, que por lo mencionado anteriormente, son datos propensos a la detección de falsos duplicados. Como

primer paso, se creó un nuevo CSV a partir del CSV unificado, para utilizar en la deduplicación. Se preservaron las columnas que sirven para la detección de autores duplicados (ORCID, nombre, apellido, filiación, CUIT, etc.), y se descartaron aquellas con información inútil para la deduplicación y que sólo aportan datos para enriquecer el perfil del autor (como la URL de la imagen o la biografía del autor). Se tomó esta decisión para facilitar el procesamiento de datos, tanto para no sobrecargar las herramientas utilizadas y disminuir los tiempos en los que procesan los datos, como para organizar las vistas de las tablas a la hora de tomar decisiones. A su vez, se agregó una nueva columna de referencia que indica la fila original del CSV unificado, lo que permite que, una vez realizada la deduplicación, se puedan obtener todos los datos de los autores desde las distintas fuentes.

Deduplicación por identificador unívoco

Este es uno de los puntos más importantes a la hora de la deduplicación, ya que el grado de confianza de los resultados obtenidos es mucho mayor que a la hora de usar otros métodos. Por este motivo, como primer caso se debe identificar cuáles datos se pueden usar como identificadores unívocos y de qué columnas se pueden obtener; en muchos de los casos se deben procesar mediante algún *script* para separar los datos que pueden servir para deduplicar, como es el caso de la columna *orcid_other_identifiers*, que trae información de otros identificadores del autor. Hecho esto, se decidió utilizar como identificador los siguientes datos: DNI, ORCID, Scopus ID, Scholar ID, Scholar URL, e-mail, Publons ID y el identificador de autoridades de SEDICI. De todas las columnas del CSV, se consideraron 8 columnas como contenedoras de identificadores unívocos de un autor:

- *sedici_authority* (columna con las claves de autoridad de los autores en SEDICI)
- *orcid*
- *scopus_id*
- *publons_id* (contenedora del Researcher ID de un autor)
- *email* (columna con el correo electrónico principal del autor)
- *emails* (una columna que contenía un listados de todos los e-mails del autor)

- *orcid_other_identifiers* (listado de otros identificadores persistentes del autor, este dato se extrajo desde ORCID por lo que podía tener el ID del autor en Google Scholar, Scopus o Publons)
- *scholar_user_id*
- *scholar_url* (columna con la URL al perfil del autor en Google Scholar).

Para las columnas *orcid*, *scopus_id*, *publons_id*, *email*, *scholar_user_url*, *scholar_user_id* y *dni*, las cuales contenían sólo un identificador por celda, se siguió la siguiente estrategia a la hora del marcado de duplicados:

1) Se abrió el CSV con la herramienta OpenRefine.
2) Se creó una nueva columna a partir de la columna que contenía el identificador usado como criterio para deduplicar, con el mismo nombre que la columna original pero se agregó el sufijo «*_filter*» al final para identificarla del original (por ejemplo, *orcid_filter* o *scopus_id_filter*). Sin embargo, a diferencia de la columna original, en la nueva columna se completaron las filas en blanco de la siguiente manera: en esta nueva columna se rellenaron las celdas vacías con el texto «vacío» seguido del número de fila en la que se encontraba la celda. Este paso es necesario ya que la herramienta detecta como duplicados a las tuplas que tienen el mismo valor, sin discriminar los valores en blanco. De esta manera se obtiene una nueva columna con valores en todas las filas lo que permite deduplicar por el identificador unívoco.
3) Luego, se ordenó a partir de esta nueva columna todo el documento y de esa manera las filas con el mismo identificador quedaron consecutivas.
4) Se marcaron como un mismo registro aquellas filas consecutivas que tenían el mismo identificador, esto se hizo utilizando una funcionalidad propia de OpenRefine sobre la nueva columna llamada «blanquear hacia abajo».
5) Por último, mediante otra funcionalidad de OpenRefine llamada «unir celdas multivaluadas», se juntaron todas las filas marcadas como un mismo registro en una misma fila y se separaron los distintos valores de una misma columna con el carácter (#).

6) Al terminar de juntar las filas, se eliminó la nueva columna, dando como resultado el CSV original pero con las filas que contenían el mismo identificador deduplicadas.

A modo de ejemplo, se muestra cómo quedaría una tabla al seguir los pasos mencionados, si se eligió como criterio para deduplicar la columna *orcid*, a partir de esta tabla:

ORCID	APELLIDO	NOMBRE	INSTITUCIÓN	ORIGINAL_ROW	FILE
0000002	Perez	Luciano Ariel	Depto. de Física	1	orcid_file.csv
0000001	Gonzales	J	Depto. de Economía	2	orcid_file.csv
	García	Tomás	UNLP	3	scopus_file.csv
0000002	Perez	Luciano	Depto. de Física	4	sedici_file.csv
0000001	Gonzales	Jorge	UNLP	5	sedici_file.csv

Al aplicar los pasos 1 y 2 este sería el resultado parcial:

ORCID	ORCID_FILTER	APELLIDO	NOMBRE	INSTITUCIÓN	ORIGINAL_ROW	FILE
0000002	0000002	Perez	Luciano Ariel	Depto. de Física	1	orcid_file.csv
0000001	0000001	Gonzales	J	Depto de Economía	2	orcid_file.csv
	'vacío'	García	Tomás	UNLP	3	scopus_file.csv
0000002	0000002	Perez	Luciano	Depto. de Física	4	sedici_file.csv
0000001	0000001	Gonzales	Jorge	UNLP	5	sedici_file.csv

Luego, al ordenar como indica el paso 3:

ORCID	ORCID_FILTER	APELLIDO	NOMBRE	INSTITUCIÓN	ORIGINAL_ROW	FILE
0000001	0000001	Gonzales	Jorge	UNLP	5	sedici_file.csv
0000001	0000001	Gonzales	J	Depto de Economía	2	orcid_file.csv
0000002	0000002	Perez	Luciano	Depto. de Física	4	sedici_file.csv
0000002	0000002	Perez	Luciano Ariel	Depto. de Física	1	orcid_file.csv
	'vacío'	García	Tomás	UNLP	3	scopus_file.csv

Con la funcionalidad de «blanquear hacia abajo» del paso 4:

ORCID	ORCID_FILTER	APELLIDO	NOMBRE	INSTITUCIÓN	ORIGINAL_ROW	FILE
0000001	0000001	Gonzales	Jorge	UNLP	5	sedici_file.csv
0000001		Gonzales	J	Depto de Economía	2	orcid_file.csv
0000002	0000002	Perez	Luciano	Depto. de Física	4	sedici_file.csv
0000002		Perez	Luciano Ariel	Depto. de Física	1	orcid_file.csv
	'vacío'	García	Tomás	UNLP	3	scopus_file.csv

Y al juntar las columnas en el paso 5:

ORCID	ORCID_FILTER	APELLIDO	NOMBRE	INSTITUCIÓN	ORIGINAL_ROW	FILE

0000001	0000001	Gonzales#Gonzales	Jorge	UNLP#Depto de Economía	2#5	orcid_file.csv#sedici_file.csv
0000002	0000002	Perez#Perez	Luciano#Luciano Ariel	Depto. de Física#Depto. de Física	1#4	orcid_file.csv#sedici_file.csv
	'vacío'	García	Tomás	UNLP	3	scopus_file.csv

Al eliminar la columna en el paso 6:

ORCID	APELLIDO	NOMBRE	INSTITUCIÓN	ORIGINAL_ROW	FILE
0000001	Gonzales#Gonzales	Jorge	UNLP#Depto de Economía	2#5	orcid_file.csv#sedici_file.csv
0000002	Perez#Perez	Luciano#Luciano Ariel	Depto. de Física#Depto. de Física	1#4	orcid_file.csv#sedici_file.csv
	García	Tomás	UNLP	3	scopus_file.csv

Como se puede observar se juntaron correctamente las filas correspondientes a los autores «Jorge Gonzales» y «Luciano Ariel Perez» en una única fila. Esto permite agrupar su información y no tenerla duplicada o esparcida en filas distintas. Además, se puede observar en la columna *original_row*, las filas del archivo original que se corresponden con las marcadas como duplicadas, y en la columna *file* se puede ver la fuente de origen de cada uno de los datos.

La secuencia de pasos descrita funcionó bien para los casos en donde las columnas contenían un solo identificador por celda, pero no hubiese funcionado si se aplicaba también a las columnas con muchos identificadores por celda, como la columna *emails* (que contenía varios mails de un autor) y *other_ids* (que contenía varios tipos de identificadores distintos); tampoco hubiera funcionado si se hubiera querido cruzar los datos entre las distintas columnas, por ejemplo, cruzar dos columnas que contengan e-mails o dos columnas que contengan el mismo

tipo de identificador, como el caso de las columnas *scopus_id*, *publons_id* y *google_scholar_url* con *other_ids*.

Para el caso de las columnas que podían tener más de un identificador en una misma celda, como el caso de la columna *other_ids*, que podía contener tanto identificadores de Scopus, Publons o Google Scholar, se realizó para cada uno de esos tipos distintos de identificadores, la siguiente serie de pasos:

1) Se armó una lista con todos los valores correspondientes al identificador persistente por el que se quería deduplicar y estos valores se extrajeron de todas las columnas que contenían al identificador. Por ejemplo, si se deduplicó por *publons_id*, la lista se armó al juntar todos los valores de la columna *publons_id* con todos los Researcher ID que se encontraban en la columna *other_ids*.
2) Luego, mediante un *script* ejecutado en OpenRefine a través de la funcionalidad «crear columna a partir de otra columna», se creó una nueva columna *identifier_filter*, e iterando sobre las celdas de la columna que contenía varios identificadores distintos, en este ejemplo *other_ids*, si algún identificador de esa columna coincidía con algún valor en la lista generada en el punto 1, entonces se agregaba, en la nueva columna *identifier_filter*, en la correspondiente fila, el índice que posee ese identificador en la lista. De esta manera, en la nueva columna, todas las filas que poseían el mismo identificador tanto en la columna *other_ids*, como en la columna correspondiente al identificador por el que se está deduplicando, quedaron con el mismo número.

Ejemplo: Si se usa *scopus_id* para deduplicar y a partir del punto 1 la lista de ID obtenida fue la siguiente:

scopus_id_list=[00001, 00002, 00005]

y se tiene el CSV con los siguientes datos:

SCOPUS_ID	OTHER_IDS
	SCOPUS_ID:00005#RESEARCHER_ID:1234

00001	
00002	
	SCOPUS_ID:00002#RESEARCHER_ID:1241#GOOGLE_SCHOLAR_URL:https://scholar.google.com/citations?user=WJp137kAAAAJ&hl=es&oi=sra

Se ve que el ID de Scopus 00001 en la segunda fila ocupa el índice 0 de la lista, el ID 00002 el índice 1 y el ID 00005 presente en la primera fila en la columna *other_ids*, el índice 2, entonces, luego de aplicar el paso 2, el CSV quedaría así:

SCOPUS_ID	OTHER_IDS	IDENTIFIER_FILTER
	SCOPUS_ID:00005#RESEARCHER_ID:1234	2
00001		0
00002		1
	SCOPUS_ID:00002#RESEARCHER_ID:1241#GOOGLE_SCHOLAR_URL:https://scholar.google.com/citations?user=WJp137kAAAAJ&hl=es&oi=sra	1

Como se puede observar, todas las filas que poseían el mismo *scopus_id*, quedaron con el mismo número en la nueva columna.

Luego de realizar este procedimiento, se deduplicaron las filas que contenían los mismos identificadores a partir de la nueva columna generada, utilizando la misma técnica de ordenar y «blanquear hacia abajo» que se usó para los casos de las columnas que contenían sólo un identificador por celda (ORCID, e-mail, etc.).

Como consecuencia de la deduplicación por identificador unívoco se marcaron como repetidos:

- 248 autores en base a su ORCID
- 1 autor en base a su Scopus ID
- 4627 autores en base a su email
- 208 autores en base a su clave de autoridad de SEDICI
- 117 autores en base a su ID en Google Scholar
- 313 en base a su DNI

La cantidad de filas se redujo de 57504 a 52878 gracias a esta deduplicación.

Deduplicación a partir del nombre completo, filiación y datos complementarios

Después de la deduplicación por identificador unívoco, el siguiente paso fue el deduplicado por el nombre completo del autor. Este tipo de deduplicación es especialmente difícil a causa de la ambigüedad propia de los nombres a la hora de identificar unívocamente a una persona: como se mencionó con anterioridad, se pueden dar casos de homónimos, segundos nombres que existan en algunas fuentes de datos pero no en otras, o la existencia de siglas en reemplazo del nombre completo. Estas características propias de los nombres hacen, por un lado, que no sea conveniente que la deduplicación a partir de ellos se realice como la hecha mediante un identificador unívoco y, por el otro, que al momento de marcar dos autores como el mismo autor utilizando como punto de referencia el nombre, se necesiten datos complementarios sobre el autor, como su lugar de trabajo, su título profesional, su dirección o datos similares que ayuden a identificar a los autores como la misma persona (sin contar los identificadores unívocos por los cuales ya se realizó la deduplicación en el paso anterior).

En la siguiente tabla se puede ver un un ejemplo de tres personas con nombres iguales o similares, la deduplicación con únicamente este dato, es propensa a obtener falsos positivos; es muy probable que Juan Pablo Gonzalez del Depto. de Física no sea la misma persona que el del Departamento de Economía y que los dos autores que pertenecen al mismo departamento sean el mismo autor.

Apellido	Nombre	Institución
Gonzalez	Juan P.	UNLP#Depto de Economía
Gonzalez	Juan Pablo	UNLP#Depto de Economía
Gonzalez	Juan Pablo	Depto. de Física#Depto. de Física

Para poder complementar al nombre en la tarea de marcar autores como repetidos, el dato elegido debía ser uno que tuvieran la mayoría de los autores (si parte de los autores no tuviera el dato, no se los podría deduplicar), y también un dato capaz de ser procesado automáticamente (no sirve, en este caso, datos como la biografía, que contiene texto que no sigue una estructura definida o algún orden). De todos los datos recopilados, uno de los que cumplía estas características era la filiación del autor. Cuando se analizan las filiaciones, se debe tener en cuenta que a medida que se comparan más adentro en el árbol de dependencia institucional (se indica como raíz del árbol de dependencias a la UNLP), las probabilidades de que haya personas con el mismo nombre en una misma filiación disminuye; es decir, es menos probable que dos personas con el mismo nombre trabajen en el mismo departamento o laboratorio y las probabilidades aumentan si se habla de facultad o incluso la UNLP en general (ver figura 24). Dos registros con el mismo nombre de autor y el mismo nombre de departamento como filiación es muy probable que se refieran al mismo registro. Sin embargo, el nombre de una filiación a veces variaba ligeramente entre las fuentes de datos, por ejemplo, el «Departamento de Letras» de la Facultad de Humanidades, a veces se encontraba como «Depto. de Letras» o «Departamento de Letras, Facultad de Humanidades». Estas variantes en el nombre dificultan el procesamiento automático y podrían haber causado confusiones a la hora de la deduplicación. Es por eso que antes de realizar la deduplicación por nombre se realizó una normalización de las filiaciones y agrupamiento de variantes.

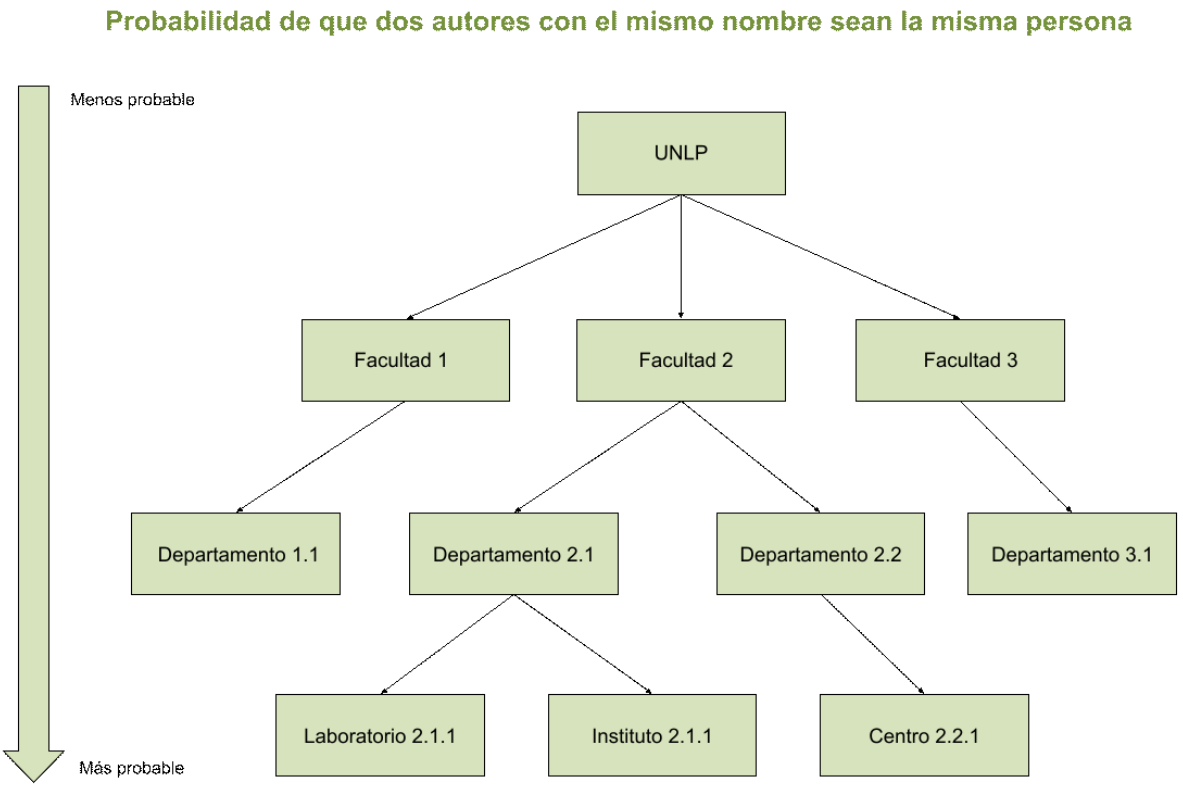


Figura 24. Gráfico que indica si es más o menos probable que dos autores con el mismo nombre sean la misma persona al tener en cuenta la institución.

Como primer paso de la normalización de las filiaciones, se obtuvieron desde la base de autoridades de SEDICI, todos los nombres de las filiaciones propias de la UNLP[4] (sus laboratorios, institutos, facultades y departamentos), junto con un ID de autoridad propio de cada filiación, y se los listó en un archivo CSV. Luego, se extrajo del archivo de autores todos los nombres de filiación que existían y se los agregó al CSV de las filiaciones extraídas desde SEDICI, y de esta manera quedó un único CSV con las filiaciones, donde en las primeras filas se encontraban las filiaciones extraídas desde SEDICI con su ID de autoridad, y al final las filiaciones extraídas del CSV de autores. A partir de este CSV de filiaciones, se marcaron como duplicados todos aquellos nombres de filiación que en realidad pertenecían a una misma filiación, y se agruparon todos estos duplicados bajo el ID de autoridad obtenido desde SEDICI. De esa forma quedaron juntas todas las variantes de un mismo nombre de filiación, y fue posible identificarlas con un ID unívoco (el ID de autoridad); todos aquellos nombres de filiaciones que no se marcaron como repetidos o no coincidían con el nombre de alguna filiación UNLP fueron

[4] Si bien este dato se obtuvo desde su base de autoridades, en SEDICI se puede ver reflejado en la siguiente comunidad: http://sedici.unlp.edu.ar/handle/10915/1

descartados. Este agrupamiento de las variantes de las afiliaciones se realizó con el uso de un software de deduplicación llamado Dedupe, el cual utiliza técnicas de aprendizaje automático para encontrar filas duplicadas en bases de datos a partir de las columnas que se le indiquen. Si bien la herramienta es privativa, la librería que utiliza se encuentra en acceso abierto en GitHub, y eso fue lo que se utilizó en el ámbito de esta tesina para ayudar a realizar el marcado de duplicados.

Luego de la normalización de las filiaciones y de obtener las variantes de los nombres de estas, se creó una nueva columna en el CSV de autores llamada *ids_affiliation*. Esta columna se completó con los ID de autoridad de SEDICI correspondientes con los nombres de filiación existentes en cada fila, es decir, si en una fila en la columna *affiliation* estaba el valor «Departamento de Física», y en el paso anterior se obtuvo que el ID de autoridad correspondiente con esa filiación es el 8, entonces en la columna *ids_affiliation* en esa fila se completó con el valor «8». Así, en el CSV de autores quedaron normalizados los nombres de filiaciones iguales bajo un mismo ID, y dos filiaciones con nombres distintos pero que son en realidad variantes de una misma filiación, pueden ser reconocidas más fácilmente como una misma filiación a partir de ese ID.

Una vez que se tuvieron los nombres de las filiaciones bajo identificadores, se procedió a realizar finalmente la deduplicación por nombre. Para realizar esto se utilizó el software Dedupe, el cual, como se mencionó anteriormente, marca filas duplicadas en archivos o bases de datos a partir de columnas y criterios configurables, y también permite indicar el tipo de dato que aloja cada columna por la que se duplica (si es un texto, un dato numérico, un nombre con formato «NOMBRE, APELLIDO», un conjunto de valores, etc.). Para empezar, se filtraron del CSV sólo las filas que tenían alguna filiación UNLP para poder usar el ID asignado en la normalización de filiaciones como criterio para el marcaje de filas repetidas. Luego se realizaron repetidos intentos de deduplicación mediante Dedupe: los resultados de los primeros intentos fueron insuficientes para dar por exitosa la deduplicación y reportaron una alta tasa de falsos positivos, lo cual se disminuyó a medida que se fue refinando la configuración utilizada en Dedupe.

Como primer intento, se configuraron las columnas del nombre completo del autor y el ID de filiación como criterios para la deduplicación. Esto significa que

Dedupe utilizó comparaciones entre las distintas filas de la columna del nombre completo del autor por un lado, y por el otro comparaciones entre las filas a partir del ID de filiación, y luego utilizó estos dos criterios para marcar las filas como repetidas. El resultado de esta deduplicación no fue satisfactorio, hubo casos, por ejemplo, en donde se marcaron como duplicados grupos de autores que compartían el apellido pero no el nombre, o en donde el ID de afiliación coincidía en los primeros dígitos pero no en el último; hubo un grupo particularmente grande de personas con apellido «García» y de primer nombre «María» pero con distinto segundo nombre que fueron marcadas todas como la misma persona.

Después de ese intento fallido de deduplicar, se revirtieron los cambios y se intentó agregar más datos sobre los autores para que el software pudiera utilizar más criterios al comparar. El dato agregado fue el e-mail de los autores, el cual podía servir en la deduplicación en casos donde hubieran quedado e-mails de autores de distintos proveedores en distintas filas, por ejemplo, en una fila podía estar el email «juanperez@example.com» y en otra «juanperez@example3.com»: en este tipo de casos esas dos filas quizás pertenezcan al mismo autor pero no se marcaron como repetidos en la deduplicación por e-mail (ID unívoco) al no ser exactamente iguales los valores de los e-mails. Luego de realizar la deduplicación con Dedupe habiendo agregado la nueva configuración, los resultados mejoraron pero tampoco fueron buenos. Alrededor de la mitad de las filas marcadas como duplicadas eran falsos positivos y muchos de los casos del primer intento en donde personas con distinto nombre fueron marcadas como la misma se mantuvieron.

Finalmente, luego de volver atrás con los cambios introducidos en el segundo intento de deduplicación, se añadió una nueva configuración a Dedupe para que existiera una interacción entre los campos del nombre del autor y el ID de filiación. Esta interacción hizo que estos campos, además de compararse por separado, también se comparen juntos, es decir, además de comparar si dos filas poseían un nombre de autor parecidos o si tenían algún ID de filiación en común, Dedupe dio más relevancia a aquellas filas en donde ambas cosas coincidían a la vez, en donde el nombre del autor con la unión de algunos de los ID de filiación coincidía. Una vez realizada la deduplicación con este nuevo criterio, los resultados

mejoraron considerablemente; se marcaron como duplicados casi la totalidad de los casos en donde el nombre del autor junto con la filiación se repetían y los casos erróneos en donde se marcaban como repetidos dos autores sólo porque compartían el apellido se redujeron casi en su totalidad.

Una vez que el marcado de autores repetidos con el uso de Dedupe fue satisfactorio, se procedió primero a eliminar los casos erróneos (los falsos positivos) y luego a juntar las filas duplicadas. Dedupe marca con un mismo ID, llamado cluster ID, aquellas filas que encontró como repetidas, y además a cada fila le asigna un valor de confianza, entre 0 y 1, que indica la seguridad con la que se marcó esa fila como duplicada, si ese número está cerca de 1, es más probable que el marcado no haya sido un falso positivo, mientras que si ese valor está cerca de 0 probablemente se trate de un falso positivo o al menos se debería revisar de forma manual si la fila realmente está duplicada. Entonces, para eliminar los casos erróneos primero se detectó a partir de qué valor de confianza las filas marcadas como duplicadas eran en realidad falsos positivos y esto se hizo revisando de manera manual las filas de los distintos valores de confianza. Cuando se encontró ese valor de confianza, se eliminaron los cluster ID de las filas con una confianza menor a dicho valor, como consecuencia se redujeron considerablemente los falsos positivos marcados como duplicados. Luego, se juntaron las filas marcadas como duplicadas y para eso se usó la misma serie de pasos utilizada para la deduplicación por identificador unívoco, tomando en este caso al cluster ID como el ID unívoco.

Como resultado alrededor de 4000 autores fueron deduplicados por su nombre y afiliación (quedaron 48890 filas de 52878).

Duplicados en celdas multivaluadas

La implementación de los métodos de deduplicación descriptos, en particular la utilización de la funcionalidad de «Unir celdas multivaluadas» de OpenRefine trajo consigo un problema: al juntar dos o más filas de un autor repetido, muchos valores de las dos filas se juntaron en una misma fila como valores multivaluados y esto podía llegar a ocasionar valores repetidos dentro de una misma celda. Por

ejemplo, si había dos filas distintas para el autor «Juan Perez» y se deduplicaron a una sola fila, pero en una de las filas, en la columna *email* estaba el valor «jperez@example3.com» y en la otra fila el valor de la columna *email* era «juanperez@example.com», entonces al deduplicar esos valores se juntaron en una misma celda como multivaluados, quedando de la siguiente manera: «jperez@example3.com#juanperez@example.com». Esto funcionó como forma de mantener dos valores distintos de un dato de un autor en una misma celda. Pero ¿qué pasaba si esos valores coincidían? Por ejemplo si ambos valores hubieran sido «jperez@example3.com», entonces la fila deduplicada hubiera quedado con los valores «jperez@example3.com#jperez@example3.com».

Para evitar esta repetición del mismo valor lo que se hizo tras cada deduplicación fue, mediante un *script* Python ejecutado con OpenRefine en cada una de las columnas, convertir los valores multivaluados a una lista, luego transformar esa lista en un *set*, estructura en la cual los valores repetidos se eliminan, y luego convertir el *set* de nuevo a una lista, para finalmente volver a transformar la lista en una cadena de caracteres, con sus elementos separador por el carácter (#). Tras este procedimiento, si había dos valores distintos multivaluados como «jperez@example3.com#juanperez@example.com» en una celda, estos se mantuvieron iguales, pero en cambio si los valores eran el mismo, como en el caso «jperez@example3.com#jperez@example3.com» entonces la celda solo se quedó con el valor «jperez@example3.com» y así se evitó la duplicación de valores en una misma celda.

Correcciones manuales

Tanto la deduplicación por ID unívoco como la deduplicación por nombre y filiación introdujeron autores mal deduplicados (falsos positivos) o no detectaron algunas duplicaciones (falsos negativos). En la deduplicación por ID unívoco los falsos positivos fueron, por lo general, resultados de errores humanos en la carga de los datos; por ejemplo, hubo un caso en donde tres autores distintos tenían el mismo email, el de la institución en donde trabajaban (por ejemplo, «info@sedici.unlp.edu.ar») como e-mail personal, y entonces se las marcó como duplicadas de manera errónea a partir de ese email. En la deduplicación por

nombre y filiación, a los errores humanos de carga de datos se les sumaron los posibles errores en la deduplicación, varios falsos positivos que pueden haber quedado dentro de los deduplicados, ya sea porque en una misma filiación justo había dos autores con exactamente el mismo nombre o porque Dedupe dio por alguna razón un valor de confianza alto a autores mal deduplicados. A su vez, si bien se deduplicaron muchos casos, varios pueden haber quedado afuera, en especial aquellas filas que hacían referencia al mismo autor pero, al tener el autor múltiples filiaciones, cada fila posee una filiación distinta: este caso si no se deduplicó mediante identificadores unívocos, es muy poco probable que haya sido detectado por Dedupe. Otros casos encontrados fueron personas con el mismo nombre, apellido y afiliación pero distinto Scopus ID (puede pasar que una misma persona tenga varios Scopus id), e incluso se encontró una persona con el mismo nombre y filiación pero distintos ORCID (esta persona tenía dos ORCID: había suspendido uno y usaba el otro).

Por estos casos mencionados fue necesario realizar sucesivas revisiones manuales sobre todo el documento para encontrar y corregir los errores, en especial los falsos positivos que hubieran hecho que la información ingresada en los perfiles fuera incorrecta. Lo que se buscó, en general, fueron filas que tuvieran valores multivaluados en columnas donde deberían tener solo un valor: por ejemplo, los identificadores persistentes (ORCID, Scopus ID, URL de Google Scholar, entre otros) deberían ser uno por persona, o casos en donde una misma fila tuviera dos nombres de autor completamente distintos, y no variantes sobre un mismo nombre como se supone que debería tener. A cada una de las irregularidades encontradas se las analizó de manera particular, y se decidió en cada caso que, si se trataba de un error de carga de los datos, el dato incorrecto se eliminaba, y si se trataba en realidad de distintos autores unidos en una misma fila, se dividía esa fila en varias, una por cada autor que estuviera mal deduplicado.

Entre la deduplicación por identificador unívoco, la deduplicación por nombre con la utilización de Dedupe y las correcciones manuales, se detectaron más de 8600 autores duplicados entre todas las fuentes. Esto permitió tanto evitar la generación de múltiples perfiles para una misma persona, como también agrupar

los datos de cada autor en una única fila y así tener los datos con la mayor completitud posible al juntarlos desde las distintas fuentes.

Selección de autores candidatos para perfiles

Con una base de datos ya normalizada y con los métodos de deduplicación mencionados aplicados, el siguiente paso fue el de decidir, de entre todos los autores recopilados, a cuáles se les generaría finalmente su perfil. En este punto se analizaron distintos criterios para la selección de dichos autores; uno podría haber sido la completitud del perfil, es decir, la cantidad de datos que se posee de un autor; por ejemplo, se podrían haber elegido los autores que sólo tuvieran una cierta cantidad de datos, o que tuvieran obligatoriamente algún dato específico, como los identificadores persistentes, el e-mail, la filiación o su DNI. Otro criterio posible podría haber sido seleccionar a los autores según si tuvieran o no citas en Google Scholar, esto por un lado hubiera permitido generar perfiles sólo a autores que posean publicaciones y, además, que alguna de esas publicaciones fuera lo suficientemente importante como para ser citada, pero, por otro lado, esto dejaría afuera a muchos autores que no tienen un perfil en Google Scholar pero que sí han realizado publicaciones académicas, sumado a que la cantidad de citas no es el único indicador de la relevancia de un autor, es más, muchas de las citas podrían ser incluso citas que realiza el mismo autor sobre su propio trabajo.

Luego de analizar las distintas posibilidades de criterios, se decidió que se les crearía un perfil solo a aquellos autores con una cantidad de publicaciones en SEDICI mayor a 20. Este criterio se eligió en base a dos puntos: el primero fue que, al ser SEDICI el repositorio objetivo de esta tesina, sus autores de mayor relevancia debían ser los primeros a los que se les creara un perfil; de esa manera, los investigadores que aportaron una mayor cantidad de obras al repositorio se verían «recompensados» con la creación de un perfil de autor para ellos. El segundo punto fue el de usar al perfil de autor como un incentivo para que los investigadores depositen sus obras; si se tienen en cuenta los múltiples beneficios que tiene la creación de un perfil para un autor (aumento en la visibilidad,

aumento en la cantidad de citas, centralización de sus obras e información en un solo lugar, entre otros) los autores de menor envergadura, o aquellos que por falta de tiempo o desconocimiento no publicaron todavía sus trabajos en SEDICI, quizás se vean motivados a depositar sus obras en el repositorio si desean obtener su propio perfil y saben que la única forma de conseguirlo es con el depósito de más de 20 obras de su autoría.

Para poder obtener de entre todos los autores recopilados aquellos con más de 20 publicaciones en SEDICI, se cruzaron los datos recopilados con los datos de SEDICI en base al ID de autoridad de los autores. Primero se obtuvo desde la base de datos de SEDICI un listado de todos los ID de autoridad de los autores junto con la cantidad de publicaciones de cada uno, luego, en el CSV de los autores recopilados, se les completó a los autores usando como guía el ID de autoridad, la cantidad de publicaciones de cada uno en una nueva columna. Finalmente, se filtraron aquellos autores con una cantidad de publicaciones mayor a 20: del total de autores recopilados quedaron 1703, y se creó un nuevo CSV sólo con esos autores.

Antes de volcar los datos de los autores elegidos en una base de datos y armar el perfil de autor, se debía revisar que todos los datos de esos autores fueran correctos, que no hubiera autores duplicados o información de los autores mezclada. Si bien ya se había aplicado un proceso de deduplicación sobre el listado y se realizaron algunas correcciones manuales, hay casos que pueden no haber sido detectados, errores en los datos traídos desde alguna fuente, o algún fallo en la deduplicación, y, para que los datos puedan mostrarse en un perfil de autor público, estos deben estar correctos en su totalidad, sin datos erróneos y lo más completos posible. Por ese motivo es que se debe realizar un control manual de cada uno de los datos de los autores candidatos para la creación de su perfil. El control manual de los datos de los 1703 autores candidatos es una tarea que conlleva una gran cantidad de tiempo y que excede los objetivos de este trabajo; por esta razón, para el contexto de esta tesina y el armado de los perfiles de autor en el prototipo funcional, sólo se consideraron de los 1703 autores, 137 pertenecientes a la Facultad de Ciencias Naturales y Museo o a alguna de sus filiaciones hijas. De este modo, la cantidad de autores se redujo a un número

aceptable para su revisión manual en el corto plazo, y, al pertenecer todos los autores a la misma facultad la posibilidad de que estén relacionados entre sí es mayor que si se hubieran elegido autores de filiaciones distintas, lo que permite que algunas estadísticas, como la de redes de coautores, tengan datos para mostrar.

La siguiente tabla muestra un resumen de cómo se redujo la cantidad de autores en cada etapa de la deduplicación, hasta llegar a los 137 autores finales:

Etapa de deduplicación	Cantidad de autores
Conjunto inicial	57504
Deduplicación por identificador unívoco	52878
Deduplicación a partir del nombre completo, filiación y datos complementarios	48890
Selección de autores candidatos (cantidad de publicaciones mayor a 20 en SEDICI)	1703
Filtrado de autores pertenecientes a la Facultad de Ciencias Naturales	137

Capítulo 8 - Implementación del prototipo

Introducción

Para la implementación de los perfiles de autor y los servicios en torno a ellos se realizó un prototipo funcional, para el cual se usaron como fuentes de datos los 137 autores resultantes de la recopilación de datos del capítulo anterior. Como base para la implementación del prototipo se utilizó la versión 7 del software para repositorios DSpace, aún en desarrollo, en su cuarta fase beta. La elección de este software para el desarrollo se basó en que, por un lado, es el software que utiliza el repositorio SEDICI, sobre el cual en un futuro se planea la integración de lo implementado en el prototipo; y por otro lado, el modelo de datos que ofrece la versión 7 de DSpace concuerda (o es lo suficientemente flexible para adaptarse) con el modelo de datos propuesto en esta tesina. Esto último facilitó el trabajo de desarrollo de los servicios propuestos, en comparación con lo que hubiera sido el desarrollo sobre una versión anterior de DSpace o sobre otro software cuyo modelo no se hubiera adaptado con tanta facilidad al modelo propuesto.

Para el desarrollo del prototipo se realizó, como primer paso, una clonación del código de la beta 4 de la versión 7 del software DSpace y así tener una base sobre la cual realizar el desarrollo. Esta versión beta de DSpace provee un repositorio funcional casi en su totalidad, incluyendo vistas personalizadas para los ítems correspondientes a entidades personas, pero no se incluyen servicios para las entidades personas y su modelo de datos y sus esquemas de metadatos no se encuentran acordes a los propuestos en esta tesina. Luego, se importó el CSV con los autores y se adaptaron los datos para que coincidieran con el modelo de datos y esquema de metadatos del prototipo. Finalmente, se realizó el desarrollo de los servicios en donde se realizaron extensiones y modificaciones tanto a la API REST que constituye el *back end* del software como a su *front end*.

Creación de la base de datos e importación de los autores

Con respecto a la base de datos utilizada en el prototipo, se utilizó la base de datos de SEDICI, pero no la base completa sino solo una comunidad pre-seleccionada de ítems, los de la comunidad correspondiente a las publicaciones de la Facultad de Ciencias Naturales y Museo. Es decir, se mantuvieron de SEDICI tanto los usuarios como los esquemas de metadatos, pero de los más de 100.000 ítems y de todas las comunidades y colecciones solo se seleccionaron los alrededor de 6000 pertenecientes a esa comunidad y a sus comunidades/colecciones hijas. Se eligió esta comunidad porque incluye muchas de las publicaciones de la institución de la que forman parte los autores recopilados a los que se les creará un perfil dentro del prototipo. De esta manera, se pudo relacionar a las publicaciones de esa comunidad con la mayoría de los autores, lo que permite una mayor riqueza en la exploración de los datos desde la interfaz.

La base de datos de SEDICI se encuentra bajo una versión de DSpace anterior a la utilizada en el prototipo (la versión 7), por lo que fue necesario realizar una migración de la base para que cumpla las características del modelo de la versión 7. Sin esta migración no hubiera sido posible contar con el modelo de entidades de datos de DSpace 7, que incluye entidades clave para el desarrollo del prototipo, como la entidad que representa a una persona o la que representa a una institución. Para ello DSpace cuenta, dentro de su variedad de comandos disponibles a través de la consola, con una herramienta de migración que permite actualizar los esquemas de la base de datos para que concuerden con la versión del software que se está utilizando.

Antes de hacer efectiva la importación hubo que adaptar el modelo de DSpace para que soporte tanto las relaciones entre entidades propuestas como la extensión del esquema de metadatos. Se crearon relaciones a nivel de base de datos entre las entidades *Persona*, *Organización* y *Publicación* (ver figura 12), para que, al momento de la importación, se pueda crear un enlace entre un autor, las publicaciones de ese autor y las organizaciones a las que pertenece. A su vez, fue necesario adaptar el esquema de metadatos para que pueda soportar todos los datos recopilados de los autores, ya que el esquema por defecto de DSpace no

tiene todos los metadatos necesarios. Entre otros, se agregaron metadatos para que contengan los Scopus ID de los autores, las URL a páginas web personales o a perfiles en redes sociales académicas, para palabras clave, variantes del nombre del autor y su DNI y CUIT.

Luego, llegó el turno de la importación de los datos: para poder hacer uso de la información de los autores recopilados, fue necesario importar el CSV con los autores elegidos en la base de datos utilizada por el prototipo. Para este fin, se hizo uso de la funcionalidad de importación de datos mediante un CSV que provee DSpace. Esta herramienta permite crear nuevos ítems dentro del repositorio a partir de un CSV que contenga en cada fila los metadatos del ítem que se desea crear, con un tipo de metadato distinto en cada columna (los nombres de las columnas deben ser el nombre del metadato que contiene cada columna). Una de las columnas en particular se corresponde con la colección del repositorio a la que pertenecerá el nuevo ítem, en este caso se creó en el repositorio prototipo una nueva colección llamada «Autores SEDICI» que sirvió de destino para los autores importados. Para indicar el tipo de entidad importada, se hizo uso del metadato *relationship.type* el cual indica justamente el tipo de entidad a la que se corresponde un ítem; en el caso del CSV a importar, ese metadato se completó con el valor *Person*, porque lo que se agregó al repositorio son representaciones de personas.

Para que el CSV de los autores se ajustara al formato requerido para importar, se cambiaron los nombres de los encabezados de las columnas y se agregaron las columnas necesarias como la del ID de la colección a la que pertenecerán los ítems y el metadato que referencia al tipo de entidad de cada ítem. El formato del CSV a importar, con los nombres de los metadatos como los encabezados de las columnas y las filas representando a los ítems se puede observar en la siguiente figura:

collection	relationship.type	person.givenName	person.familyName	person.nameVariant
123456789/50	Person	Francisco Raúl	Carnese	Carnese, Francisco Raúl||Francisc
123456789/50	Person	Alicia Bibiana	Orden	Orden, Alicia Bibiana
123456789/50	Person	Gustavo Alberto	Darrigran	Gustavo Darrigran||Darrigran, Gust
123456789/50	Person	Jorge Luis	Frangi	Frangi, Jorge||Frangi, Jorge L.||Fra
123456789/50	Person	Marcelo Fabián	Arturi	Arturi, Marcelo Fabián||Arturi, Marc
123456789/50	Person	Juan Francisco	Goya	Goya, J.||Goya, J. F.||Goya, Juan|

Figura 25. *Vista de la estructura del archivo CVS a importar*

Después de importar la información de los autores, se hizo lo mismo con los datos de las organizaciones pertenecientes a la UNLP. Se creó un CSV que contenía los nombres de todas las instituciones pertenecientes a la UNLP junto con su clave de autoridad de SEDICI, y luego se importó ese archivo en la base de datos del prototipo, en este caso el valor del metadato *relationship.type* no fue *Person* sino que fue *OrgUnit* para indicar que se trataba de organizaciones. De la misma manera que se creó en el prototipo una colección para alojar a los autores, también se creó la colección «Instituciones UNLP» para que sirva como contenedora de las instituciones importadas.

Como paso final en el armado de la base de datos, se procedió a relacionar los autores importados con las publicaciones e instituciones correspondientes. Las publicaciones tienen entre sus metadatos la clave de autoridad propia de SEDICI de muchos de sus autores, a su vez, los autores tienen la clave de autoridad de las instituciones a las que pertenecen, además de su propia clave de autoridad. Entonces, con el uso de las claves de autoridad de SEDICI y los metadatos de relación que ofrece DSpace 7, se forjaron enlaces entre estos tres tipos de entidades, lo que facilita la navegación entre ellas y la visualización en los perfiles de autor de las publicaciones que el autor realizó y de las instituciones a las que pertenece.

Implementación de los servicios

Para el desarrollo de los servicios propuestos en el prototipo, se realizaron extensiones y modificaciones a la beta 4 de la versión 7 del software DSpace. A diferencia de sus versiones anteriores, la versión 7 de DSpace separa en dos proyectos distintos su *back end* y su *front end*. Por un lado, el *back end* se trata de una API REST desarrollada en Java, la cual utiliza el *framework* Spring y sigue los principios HATEOAS para aplicaciones REST. El *front end*, por otro lado, es una aplicación desarrollada sobre el framework Javascript Angular, la cual se comunica con la API REST Java a través de sus distintos *end point* para obtener los datos necesarios y mostrarlos en una interfaz de usuario (UI)interfaz de usuario (UI)

amigable al usuario final. La mayoría de las modificaciones y extensiones realizadas se efectuaron sobre el *front end* Angular, dado que casi no fue necesaria la modificación de ninguna funcionalidad de la API REST; solo en el caso del servicio de estadísticas, por una cuestión de performance, se agregaron nuevos *endpoints* a la API.

De la totalidad de los servicios propuestos para el prototipo sólo se implementaron un subconjunto de ellos, los de mayor importancia dentro de los propuestos. Esta decisión se tomó como consecuencia de los acotados tiempos de desarrollo de la tesina. El resto de los servicios quedaron pendientes para su implementación en el futuro y quizás se podrían integrar en una segunda versión del prototipo.

Los servicios que efectivamente se implementaron fueron:

- La visualización pública de los perfiles de autor
- La navegación entre las relaciones de un autor
- La visualización de estadísticas sobre los datos y publicaciones de los autores
- La posibilidad de exportar el perfil de autor en formato PDF
- La generación de un código QR que contiene un enlace al perfil del autor
- La búsqueda y visualización de resultados de un autor dentro del repositorio a partir de sus datos (nombre, identificador persistente, etc.)

Los servicios cuya implementación quedó pendiente a futuro fueron:

- La exposición del perfil bajo algún protocolo de interoperación (RDF, OAI, etc.)
- La posibilidad de que un autor pueda loguearse al repositorio como tal y allí modificar su perfil y sus datos de investigador
- La posibilidad de conectar el perfil de autor con su equivalente en ORCID e intercambiar datos enriqueciendo la información que ofrece el perfil a partir de ello
- La petición de corrección de datos erróneos en un perfil de autor por parte de cualquier usuario, entre otros

Las extensiones realizadas tanto en el lado del *back end* como del *front end* intentaron adaptarse a la estructura de clases y componentes existente dentro de los respectivos proyectos. En el *back end* se continuó con la estructura propuesta por Spring de servicios y componentes que realizan la conversión de los datos enviados dentro de una *request* a objetos del modelo de datos y viceversa. Mientras que en el *front end* desarrollado en Angular, se agregaron nuevos componentes[5] que permitan la visualización de los nuevos servicios. Los nuevos componentes se situaron dentro de un componente preexistente, encargado de mostrar los datos de una persona, para así conformar el prototipo del perfil de autor, el componente *person-page*. A continuación se explicará en detalle la implementación de cada uno de los servicios que finalmente formaron parte del prototipo.

Visualización pública del perfil de autor: sus componentes, relaciones y la navegación entre ellas

El *front end* Angular de la versión 7 de DSpace contiene entre sus funcionalidades la visualización en una misma página de algunos metadatos de una entidad-persona junto con sus publicaciones y organizaciones y a las que pertenece, o en otras palabras, un perfil de autor que sólo se limita a exponer un subconjunto del total de los datos de un autor junto con sus publicaciones y relaciones. Además, en el proyecto existen varios componentes que permiten la exposición del resto de los metadatos de un autor pero que no son usados en esa versión de DSpace en particular. A su vez, en DSpace 7 ya se permite la visualización pública de este perfil de autor, es decir, a cada una de estas páginas que exponen la información de una entidad-persona se le otorga una URL propia dentro de la instancia del software, a la que no es necesario estar autenticado dentro del sistema para acceder.

[5] En Angular un componente es un bloque de código reutilizable, que representa una sección de la UI.

Todas estas funcionalidades preexistentes en el proyecto agilizaron la implementación propia de algunos de los servicios propuestos en el prototipo y otorgaron una base sobre la cual completar la implementación del perfil de autor y el resto de los servicios. Esto implicó que, por ejemplo, no hiciera falta modificar el proyecto para agregar una página propia de perfil de autor, ni para la visualización de muchos de los datos del autor y alguna de sus relaciones y la navegación entre ellas, o incluso para que esta página se visualizara sin ningún tipo de autenticación. Pero para mostrar la totalidad de los metadatos de un autor propuestos en el armado del prototipo, sí hubo que modificar el componente encargado de mostrar los datos del autor, el componente *person-page*. Estas modificaciones comprendieron:

- La inclusión de la visualización de metadatos de un autor que no se mostraban por defecto en DSpace pero que sí se propusieron en el prototipo, como los identificadores Scopus ID, Google Scholar ID y Researcher ID, URL a páginas web o blogs propios del autor, el e-mail y el número de teléfono de un autor y las palabras clave que describen su campo de estudios.
- La creación e inclusión al perfil de un componente que agrupe y liste todas las palabras clave que describen el campo de estudio de un autor, y que al hacer clic sobre cualquiera de ellas se redireccione a una búsqueda de todos los ítems del repositorio que se relacionen con esa palabra clave.
- La creación e inclusión de un nuevo componente que agrupe todos los identificadores persistentes de un autor bajo una misma sección de la pantalla de perfil de autor.
- La inclusión de nuevos componentes que permitan visualizar los servicios implementados en el prototipo cuya implementación se describe en posteriores secciones de este capítulo de la tesina, por ejemplo, el código QR del perfil de autor.
- La inclusión de un menú de solapas para permitir al usuario elegir si desea ver las publicaciones de un autor o el reporte de estadísticas, por debajo de la visualización datos del autor.

- El reordenamiento de todos los elementos que conforman el perfil de autor dentro del componente *person-page*, de manera tal que se muestren en tres columnas distintas de la pantalla para visualización más ordenada.

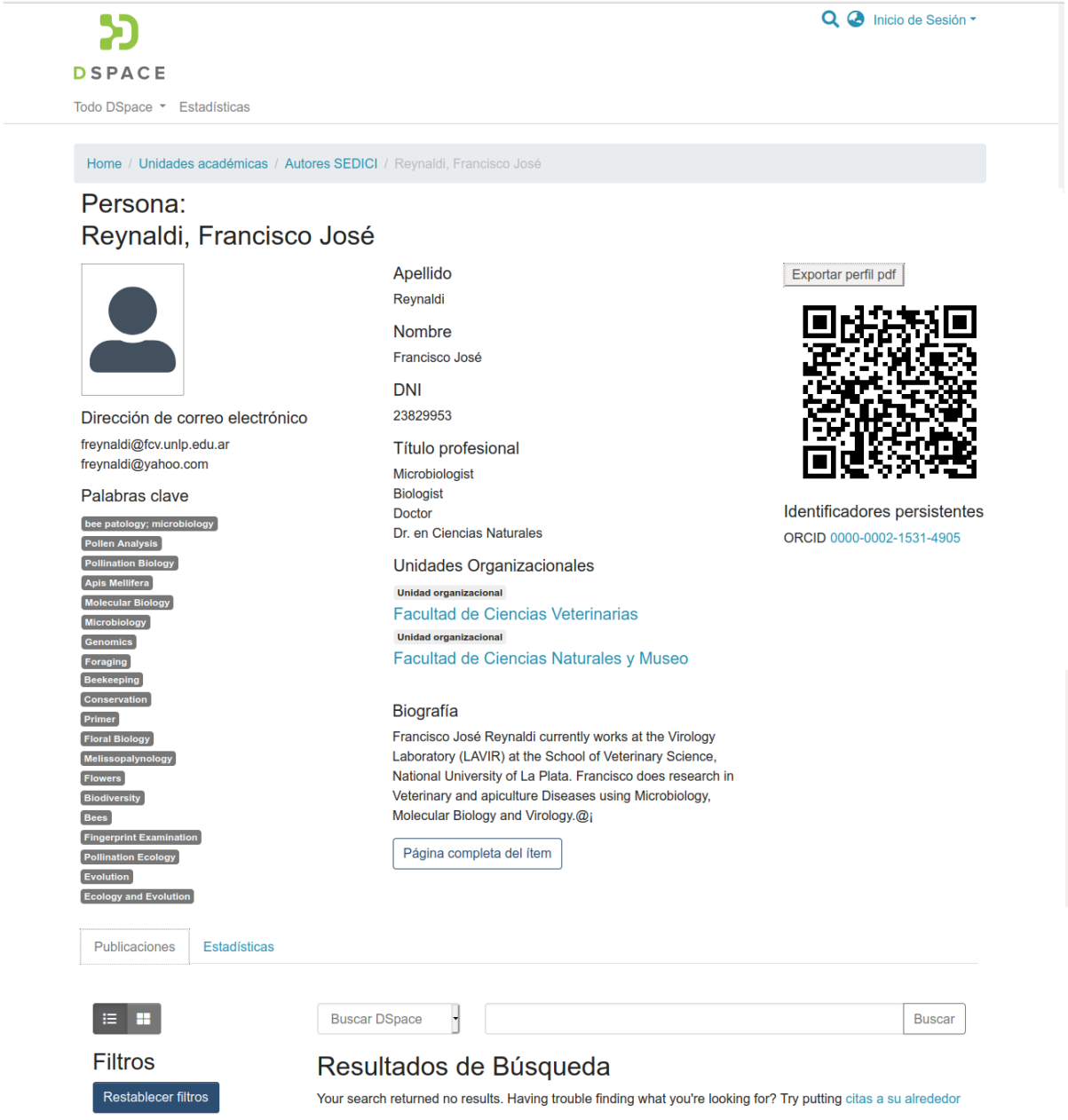


Figura 26. *Ejemplo de perfil de autor implementado en el prototipo*

La figura 26 muestra un ejemplo de perfil de autor resultante luego de las modificaciones descritas. Se puede observar cómo los datos del autor están separados en tres columnas y por debajo de ellos el menú de pestañas con las opciones de publicaciones y estadísticas. En la esquina superior derecha de la

imagen se puede observar que no se está autenticado dentro del sistema, lo que implica que el perfil se muestra de manera indistinta tanto para los usuarios que tienen un usuario como para los que no. Por debajo de la imagen y de los correos electrónicos del perfil, se encuentra el listado de palabras claves; cada palabra clave se encuentra remarcada y al hacer clic en alguna de ellas se permite la exploración de los recursos del repositorio que contengan esa palabra clave. En el centro del perfil se listan las organizaciones a las que pertenece el autor, y se puede ir a la página del repositorio propia de cada organización si se selecciona cualquiera de ellas, y por debajo de ellas se encuentra una breve biografía. En la columna derecha del perfil se puede ver el listado de identificadores persistentes y páginas web pertenecientes al autor (en este ejemplo el autor no posee páginas web), justo por debajo del código QR y del botón que permite exportar el perfil en formato PDF. Por debajo de las tres columnas de datos personales del autor aparece un menú de solapas que da la opción de visualizar o bien el listado de las publicaciones de ese autor o bien sus estadísticas.

Visualización de estadísticas

Para el desarrollo del prototipo se incluyeron estadísticas de la cantidad de publicaciones que realizó un autor a lo largo del tiempo, la cantidad de publicaciones de un autor agrupadas por tipo de publicación, la red de coautores del autor y, además, dos estadísticas de uso como la cantidad de visualizaciones que posee el perfil de un autor y la cantidad de visualizaciones al perfil agrupadas por año. La implementación de los primeros tres reportes de estadísticas mencionados requirió cambios y extensiones tanto en la API REST *back end* del proyecto como en la UI desarrollada en Angular, mientras que para los reportes de las estadísticas de uso solo fue necesario realizar modificaciones en el *front end*.

La modificación del *back end* para el desarrollo de la visualización de estadísticas fue consecuencia de que en DSpace 7 no existen *endpoints* para las estadísticas de las publicaciones de un autor, por lo que fue necesario agregarlos para poder implementar ese servicio. Si bien se podrían haber generado todos los reportes directamente desde el *front end* de Angular a partir de un *endpoint* que

retorna todas las publicaciones de un autor dado, esto hubiera requerido el procesamiento de una cantidad considerable de datos por parte del cliente, lo que hubiera derivado en una disminución de los tiempos de respuesta del sistema. El reporte de las estadísticas de uso, por su parte, ya tenía asociado *endpoints* en el *back end* por lo que sólo fue necesario realizar las modificaciones correspondientes en la UI para poder visualizarlo dentro de la página de perfil de autor.

Cambios realizados a la API REST

Los cambios realizados al *back end* trataron de seguir la estructura existente de DSpace 7, en donde la arquitectura responsable de definir los *endpoints* de la API REST y procesar una *request* está conformada, en una versión simplificada, por las clases de la siguiente figura:

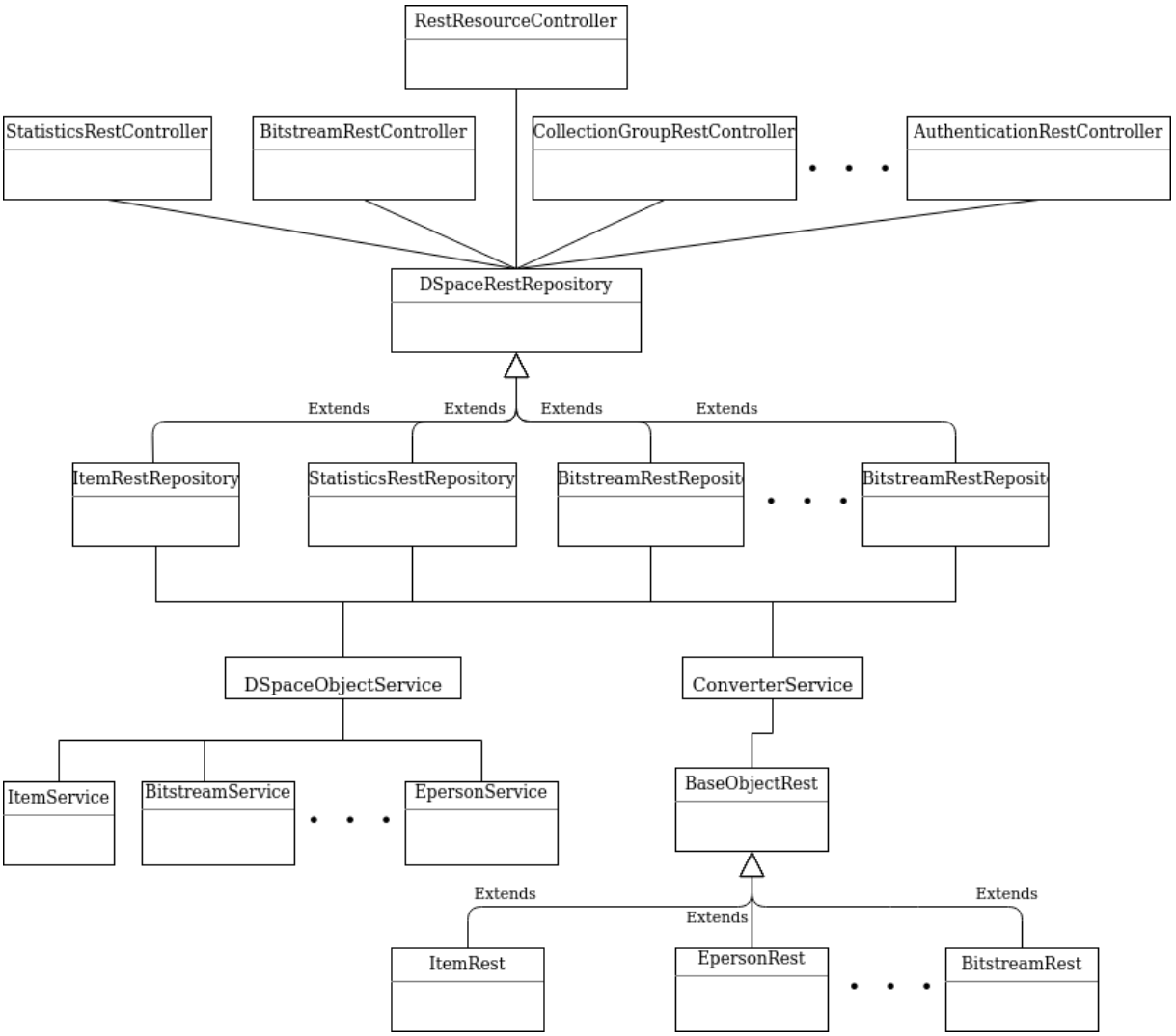


Figura 27. *Diagrama de clases de la API REST de DSpace 7*

El punto de entrada para la API REST es la clase *RestResourceController*: allí se definen de manera genérica todos los *endpoints* de la aplicación. Esta clase funciona como «router» de la aplicación: dependiendo del tipo de *request* y el modelo de la aplicación objeto de la *request*, delega la lógica de resolución de la *request* a la clase hija de *DSpaceRestRespository* que corresponda. Para los *endpoints* específicos de un modelo en particular, *RestResourceController* delega su resolución a los controladores propios de cada modelo para que estos los resuelvan por su cuenta, y luego estos dejan la lógica de la acción a realizar en manos del correspondiente *RestRepository*. Luego, la clase *RestRespository* encargada de resolver la petición por lo general realiza dos acciones: primero, realiza la acción requerida mediante alguno de los servicios disponibles y obtiene el resultado de dicha acción. Luego, transforma el resultado en una respuesta que

siga las convenciones REST impuestas por HATEOAS. Esto lo logra a partir de un converter (convertidor), el cual toma los datos del modelo de base de datos, devueltos por los servicios y los transforma a su correspondiente modelo REST. Por ejemplo, si el dato retornado por el servicio se trata de un *DspaceItem*, entonces el converter transformará ese ítem en un *ItemRest*. Finalmente, el correspondiente *RestRepository* devolverá la respuesta a los controladores para que estos completen la resolución de la *request*.

Para la implementación del reporte estadísticas de un autor en el prototipo, a esta estructura se le añadieron tres nuevos *endpoints*, uno para el reporte del número de publicaciones de un autor agrupados por tipo, otro para el reporte de la cantidad de publicaciones a lo largo de los años, y el último para la red de coautores de un autor. Estos *endpoints* fueron definidos dentro de unos de los controladores ya existentes, el *StatisticsRestController*. A su vez, se creó otra clase del tipo *RestRepository*, el *PersonStatisticsRestRepository*, encargado de efectuar la lógica de resolución de cada uno de los tres *endpoints*, es decir, por cada *endpoint* definido en el *StatisticsRestController*, se hace un llamado a un método de *PersonStatisticsRestRepository* que resuelve ese *request*. De la misma manera, esta última clase utiliza un servicio para que realice las acciones necesarias que retornen los datos esperados. Para eso, se creó un servicio llamado *PersonStatisticsProvider*, el cual retorna, a través de llamados a otros servicios y el procesamiento de los datos devueltos por ellos, la información requerida para cada uno de los tres reportes de estadísticas distintos. Finalmente, luego de obtener los datos del servicio, la clase *PersonStatisticsRestRepository* convierte los datos a un formato REST y los retorna al controlador para que devuelva el reporte esperado. El formato REST de cada uno de los tres reportes se definió en la clase *PersonStatisticsRest*.

Se puede visualizar cómo las extensiones realizadas siguen el formato de clases ya existente en DSpace 7 descripto anteriormente en la siguiente figura:

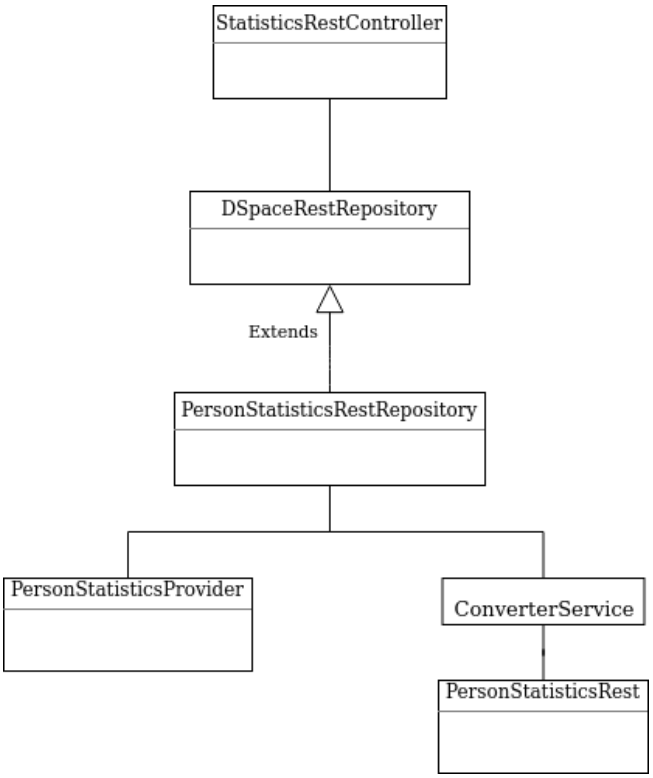


Figura 28. *Clases creadas y modificadas para el prototipo en la API REST de DSpace*

Cambios realizados al *front end* Angular

Para introducir el servicio de estadísticas de un autor en la UI, se añadió un nuevo componente llamado *person-statistics*, el cual representa la sección dentro de la pantalla en donde se podrán visualizar las estadísticas de un autor. Este componente se situó dentro del componente ya existente *person-page*, que muestra todos los datos de una persona dada, dentro de un menú en forma de tabs propio del componente *person-page*. En este menú, además del componente de estadísticas se agregó el componente que muestra las publicaciones de un autor para que el usuario o bien vea las publicaciones o bien las estadísticas de un autor, dependiendo de la opción que elija.

A su vez, dentro del componente de estadísticas, se crearon tres componentes más, cada uno correspondiente con los nuevos tipos de estadísticas creados: un componente para visualizar las publicaciones de un autor agrupadas por tipo, otro para mostrar las publicaciones de un autor a lo largo del tiempo y otro que permite visualizar la red de coautores del autor. Y, además de estos tres nuevos

componentes, también se agregó un componente de visualización de estadísticas de uso, preexistente en el proyecto, el cual permite ver la cantidad de visitas que tiene la página de un determinado ítem de DSpace (en este caso, el ítem es la entidad persona a la que hace referencia el perfil). Todos estos componentes toman datos provistos como entrada y los transforman generando gráficos que permiten visualizar de forma amena para el usuario la información a reportar. Para la generación de estos gráficos se utilizó la librería ngx-echarts, la cual permite la generación de gráficos de barra, de torta, grafos, árboles e incluso gráficos 3D e interactivos a partir datos provistos en simples estructuras de datos y pequeñas modificaciones a los componentes en donde se los quiera visualizar.

Las siguientes son imágenes de la versión final del reporte de estadísticas en el prototipo, junto con el menú de solapas añadido en el componente *person-page*. Todas las imágenes pertenecen a la misma página del prototipo, correspondiente al perfil del autor «Salceda, Susana Alicia»:

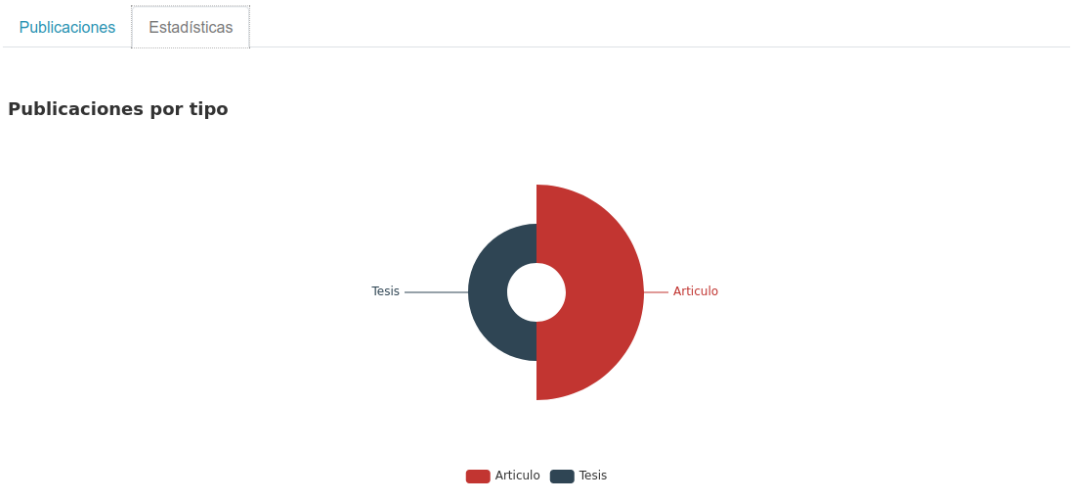


Figura 29. *Menú en forma de solapas con las opciones «publicaciones» y «estadísticas» junto con el gráfico de publicaciones agrupadas por tipo*

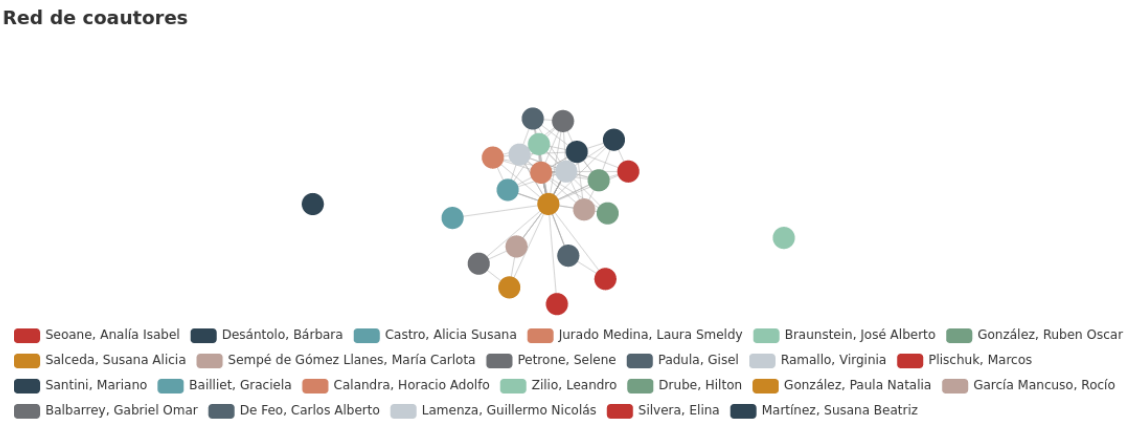


Figura 30. *Red de coautores de un autor*

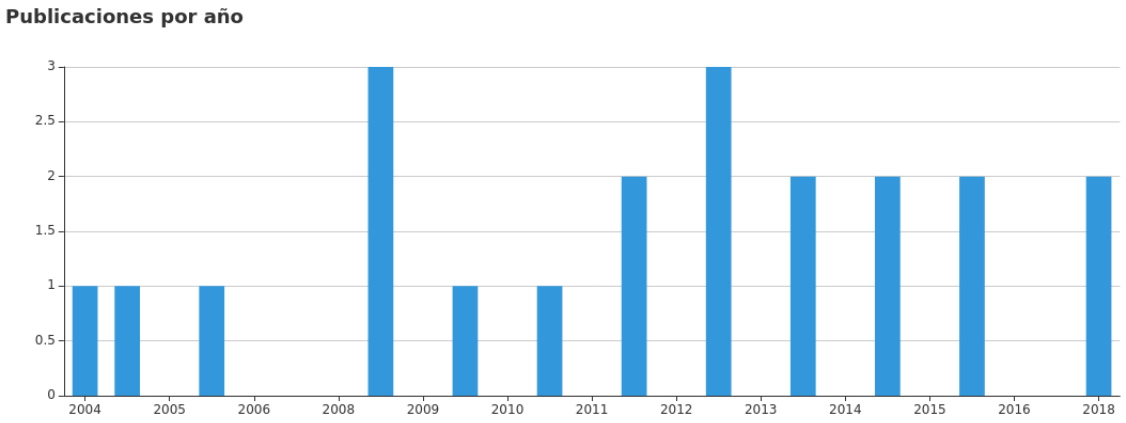


Figura 31. *Gráfico de las publicaciones por año de un autor*

Total visits

	views
Salceda, Susana Alicia	0

Total visits per month

	views
August 2020	0
September 2020	0
October 2020	0
November 2020	0

Figura 32. *Estadísticas de uso de la página de perfil de autor*

Para proveer de los datos necesarios a los tres nuevos componentes, se creó un nuevo servicio Angular. Estos servicios son los encargados de obtener los datos

que luego serán mostrados al usuario en los componentes; estos datos pueden obtenerse a partir de *requests* a API REST o a algún otro servicio que funcione como *back end*. En este caso el servicio creado obtiene los datos a partir de *requests* HTTP a los tres *endpoints* que se crearon en la API REST del proyecto para este prototipo, y que se describen en el capítulo anterior. En la siguiente figura se puede observar gráficamente la interacción entre el servicio con la API REST y luego con los tres nuevos componentes.

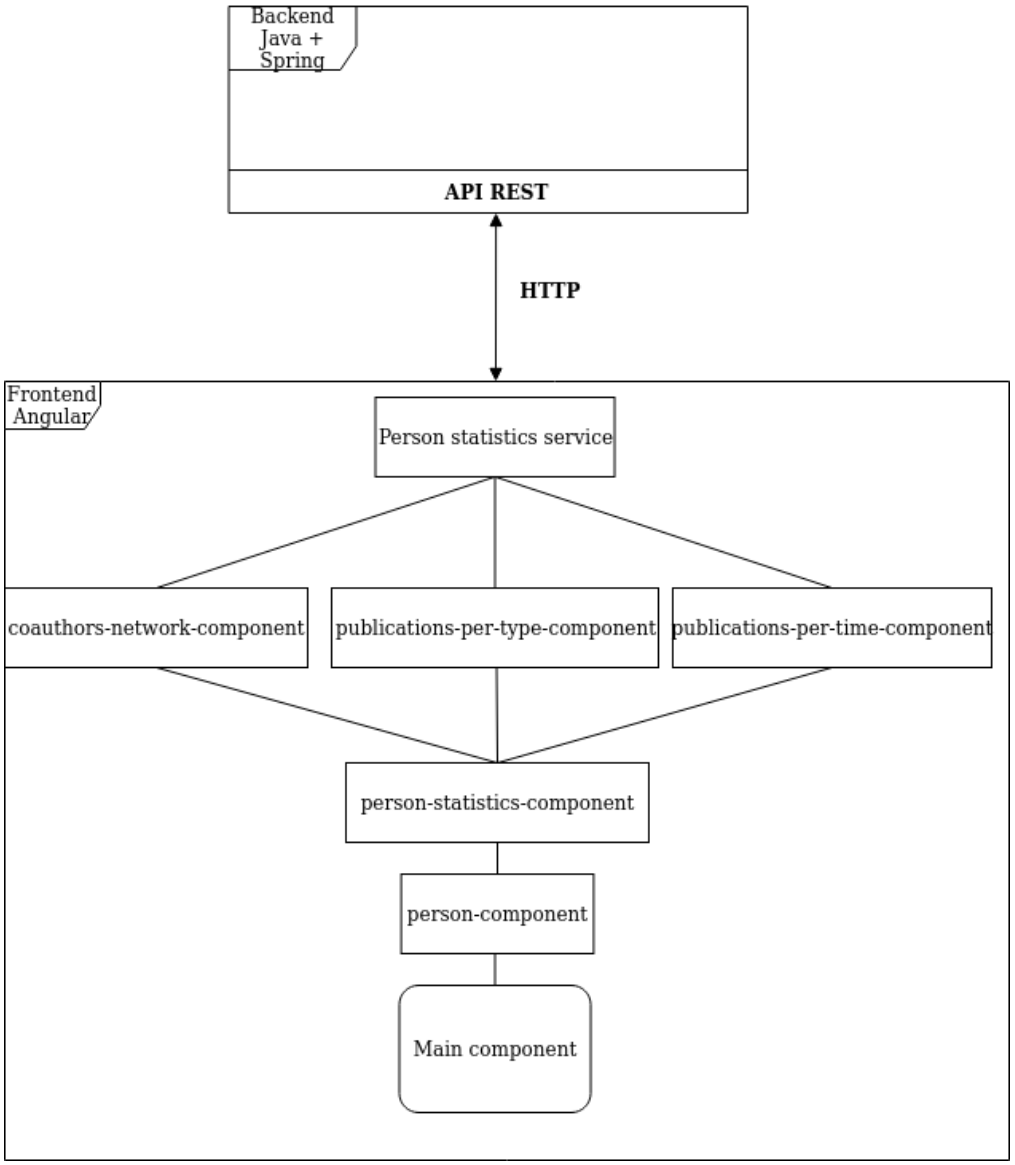
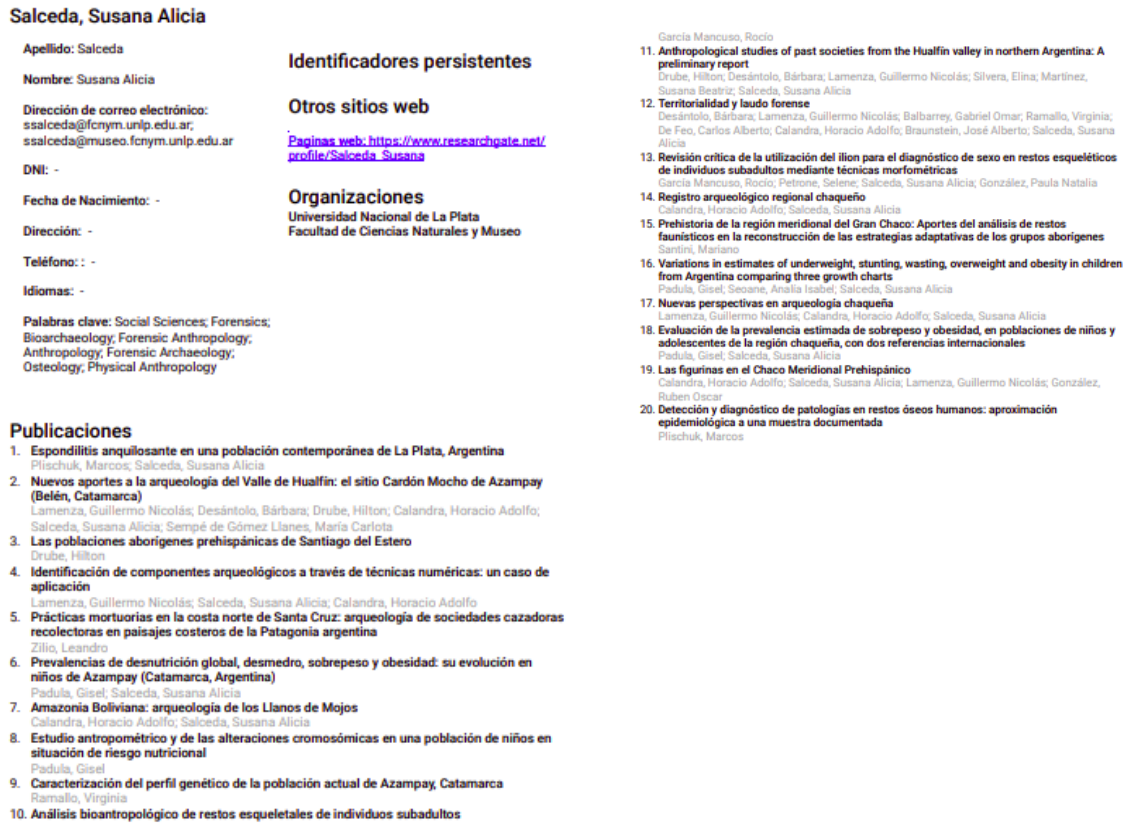


Figura 33. *Interacción de los servicios y componentes creados en el front end Angular del prototipo*

Exportación del perfil de autor

Uno de los servicios planteados para realizar en este prototipo es el de realizar un perfil que permita exportar mediante algún formato los datos del autor como si fuese una especie de *curriculum vitae*. Esto puede ser útil para el investigador a la hora de querer compartir las publicaciones y datos personales de forma electrónica o impresa. Para implementar este servicio se generó un nuevo componente, *person-export-profile*, que se integra en el componente *person-page* y mediante la librería pdfmake de Angular se logra exportar los datos del autor. En la siguiente figura se puede ver una captura del PDF generado al exportar el perfil de un autor.



Salceda, Susana Alicia

Apellido: Salceda

Nombre: Susana Alicia

Dirección de correo electrónico: ssalceda@fcnym.unlp.edu.ar; ssalceda@museo.fcnym.unlp.edu.ar

DNI: -

Fecha de Nacimiento: -

Dirección: -

Teléfono: : -

Idiomas: -

Palabras clave: Social Sciences; Forensics; Bioarchaeology; Forensic Anthropology; Anthropology; Forensic Archaeology; Osteology; Physical Anthropology

Identificadores persistentes

Otros sitios web

Paginas web: https://www.researchgate.net/profile/Salceda_Susana

Organizaciones

Universidad Nacional de La Plata
Facultad de Ciencias Naturales y Museo

Publicaciones

1. Espondilitis anquilosante en una población contemporánea de La Plata, Argentina
   Plischuk, Marcos; Salceda, Susana Alicia
2. Nuevos aportes a la arqueología del Valle de Hualfín: el sitio Cardón Mocho de Azampay (Belén, Catamarca)
   Lamenza, Guillermo Nicolás; Desántolo, Bárbara; Drube, Hilton; Calandra, Horacio Adolfo; Salceda, Susana Alicia; Sempé de Gómez Llanes, María Carlota
3. Las poblaciones aborígenes prehispánicas de Santiago del Estero
   Drube, Hilton
4. Identificación de componentes arqueológicos a través de técnicas numéricas: un caso de aplicación
   Lamenza, Guillermo Nicolás; Salceda, Susana Alicia; Calandra, Horacio Adolfo
5. Prácticas mortuorias en la costa norte de Santa Cruz: arqueología de sociedades cazadoras recolectoras en paisajes costeros de la Patagonia argentina
   Zilio, Leandro
6. Prevalencias de desnutrición global, desmedro, sobrepeso y obesidad: su evolución en niños de Azampay (Catamarca, Argentina)
   Padula, Gisel; Salceda, Susana Alicia
7. Amazonia Boliviana: arqueología de los Llanos de Mojos
   Calandra, Horacio Adolfo; Salceda, Susana Alicia
8. Estudio antropométrico y de las alteraciones cromosómicas en una población de niños en situación de riesgo nutricional
   Padula, Gisel
9. Caracterización del perfil genético de la población actual de Azampay, Catamarca
   Ramallo, Virginia
10. Análisis bioantropológico de restos esqueletales de individuos subadultos
    García Mancuso, Rocío
11. Anthropological studies of past societies from the Hualfín valley in northern Argentina: A preliminary report
    Drube, Hilton; Desántolo, Bárbara; Lamenza, Guillermo Nicolás; Silvera, Elina; Martínez, Susana Beatriz; Salceda, Susana Alicia
12. Territorialidad y laudo forense
    Desántolo, Bárbara; Lamenza, Guillermo Nicolás; Balbarrey, Gabriel Omar; Ramallo, Virginia; De Feo, Carlos Alberto; Calandra, Horacio Adolfo; Braunstein, José Alberto; Salceda, Susana Alicia
13. Revisión crítica de la utilización del ilion para el diagnóstico de sexo en restos esqueléticos de individuos subadultos mediante técnicas morfométricas
    García Mancuso, Rocío; Petrone, Selene; Salceda, Susana Alicia; González, Paula Natalia
14. Registro arqueológico regional chaqueño
    Calandra, Horacio Adolfo; Salceda, Susana Alicia
15. Prehistoria de la región meridional del Gran Chaco: Aportes del análisis de restos faunísticos en la reconstrucción de las estrategias adaptativas de los grupos aborígenes
    Santini, Mariano
16. Variations in estimates of underweight, stunting, wasting, overweight and obesity in children from Argentina comparing three growth charts
    Padula, Gisel; Seoane, Analía Isabel; Salceda, Susana Alicia
17. Nuevas perspectivas en arqueología chaqueña
    Lamenza, Guillermo Nicolás; Calandra, Horacio Adolfo; Salceda, Susana Alicia
18. Evaluación de la prevalencia estimada de sobrepeso y obesidad, en poblaciones de niños y adolescentes de la región chaqueña, con dos referencias internacionales
    Padula, Gisel; Salceda, Susana Alicia
19. Las figurinas en el Chaco Meridional Prehispánico
    Calandra, Horacio Adolfo; Salceda, Susana Alicia; Lamenza, Guillermo Nicolás; González, Rubén Oscar
20. Detección y diagnóstico de patologías en restos óseos humanos: aproximación epidemiológica a una muestra documentada
    Plischuk, Marcos

Figura 34. *Exportación del perfil de autor a PDF*

Código QR del perfil

Uno de los servicios propuestos para el prototipo fue la inclusión de un código QR, el cual, al escanearlo, redireccione hacia la misma página del perfil de autor. Esto puede llegar a ser útil al momento de que un autor realice presentaciones en algún congreso y quiera obtener mayor visibilidad al compartir su perfil con el resto de la comunidad; allí el autor podría adjuntar el código QR de su perfil dentro de su presentación o póster. Para lograr el armado del QR propuesto se creó un nuevo componente, el cual obtiene la URL de la página en la que se encuentra a partir de una función JavaScript y luego, con el uso de la librería angularx-qrcode, crea el código QR a partir de la URL y lo muestra en pantalla. Este nuevo componente se incluye en el componente *person-page* para que el código se muestre dentro de la página de perfil de autor.

Una característica interesante de este nuevo componente es que no sólo se puede utilizar para las páginas de perfil de autor, sino que se puede introducir en cualquier página que se quiera. Para lograr esto, solamente es necesario incluir el nuevo componente que genera el código QR en el componente encargado de mostrar una página particular y entonces el QR se creará a partir de la URL de esa página. De esta manera se puede tener códigos QR para la página principal de un repositorio, para una publicación o incluso para la página de una organización dentro del repositorio. En la figura 26 se puede observar un ejemplo del código QR aplicado en un perfil de autor.

Búsqueda de un autor a partir de sus datos

La versión 7 de DSpace en su beta 4 trae entre sus características la posibilidad de realizar búsquedas a partir de valores presentes en cualquiera de los metadatos de un ítem. En otras palabras, si se realiza una búsqueda por la palabra ‘naturales’, entonces el motor de búsqueda incluirá entre los resultados todos los ítems que tengan en cualquiera de sus metadatos el valor ‘naturales’. Como consecuencia de esta característica es posible la búsqueda de un autor a partir de cualquiera de sus datos desde la página de búsqueda de DSpace 7, dado que al ser

un autor (es decir una entidad-persona) es un ítem más dentro del repositorio. Por eso, lo único que se modificó de la beta 4 de DSpace 7 para que se permita la búsqueda de un autor a partir de cualquiera de sus datos, fue el agregado de los metadatos que forman parte del prototipo y que no existen por defecto en DSpace. Con la simple inclusión de estos metadatos y otorgándole valores ya es posible la búsqueda de un autor por su Scopus ID o sus palabras clave. Este agregado de los nuevos metadatos fue realizado en conjunto con el llenado de la base de datos y se explica en el capítulo de creación de la base de datos e importación de los autores de esta tesina.

Un ejemplo de esta funcionalidad de búsqueda aplicada al prototipo se puede ver en la figura 35, en donde se buscó a partir del valor '0000-0001-5793-8882' el cual coincide con el ORCID de la autora «Ana Sabrina Mora».

Figura 35. *Búsqueda en el prototipo a partir de un ORCID*

Capítulo 9 - Conclusiones

A raíz del estudio realizado sobre perfiles de autor y su implementación en repositorios institucionales se llegó a la conclusión de que el desarrollo de estos perfiles para un repositorio como el de la UNLP, no sólo aumentaría la visibilidad del trabajo de los autores y del repositorio en sí mismo, sino también que abriría la puerta al desarrollo de servicios de valor agregado alrededor de estos perfiles de autor. Estos servicios, a su vez, servirían como incentivo para que los autores depositen más obras en el repositorio, lo que atraería una mayor cantidad de autores que quieran formar parte de él.

Se implementó un prototipo funcional de un repositorio, basado en la versión 7 del software DSpace, al cual se le realizaron modificaciones para que fuera posible la visualización pública de perfiles de autor junto con un conjunto de servicios en torno a ellos. Para poblar al prototipo con los datos de autores reales se realizó una recopilación de datos de autores UNLP desde diversas fuentes. Cada una de estas fuentes contenía un conjunto de datos de autores que variaba de fuente en fuente, es por eso que se debió realizar un proceso de unión y normalización de los datos, seguido de un proceso de deduplicación para los casos en los que un mismo autor poseía su información dispersa en las distintas fuentes. Todo esto se llevó a cabo mayoritariamente con el uso de la herramienta OpenRefine y, al finalizar la recopilación, se obtuvo, desde nueve fuentes distintas, un total de 57.504 autores con sus respectivos datos, los cuales luego del proceso de normalización y deduplicación se redujeron a 48.890.

Esa gran cantidad de autores y sus datos constituyen una valiosa base de datos que podría servir de fuente para la futura implementación de perfiles de autor en SEDICI, o incluso servir para completar y aportar mayor cantidad de información a la base de autoridades de dicho repositorio. Sin embargo, esta cantidad de autores es excesiva para los propósitos de este prototipo, de modo tal que sólo se seleccionó un subconjunto de ellos a la hora de crear la base de datos del prototipo. Se seleccionaron de los 48.890 solo 137, que se corresponden a autores

pertenecientes a la Facultad de Ciencias Naturales y Museo o a alguna de sus filiaciones hijas, que tuvieran más de 20 publicaciones en SEDICI. Esta selección permitió tener una cantidad de autores manejable a la hora de realizar verificaciones manuales y, a su vez, al pertenecer todos los autores a la misma institución, era más probable que poseyeran relaciones de coautoría entre ellos, cosa que sería útil a la hora de la visualización de estadísticas o de la navegación entre las distintas entidades. Además de los autores, para conformar la base de datos del prototipo se obtuvieron todas las instituciones, organizaciones y unidades de I+D pertenecientes a la UNLP, sumados a las publicaciones en SEDICI de cada uno de estos autores.

Para el desarrollo del prototipo primero se definieron tanto el modelo de datos y el esquema de metadatos, como los servicios que formarían parte de él. Como base para la implementación, se eligió la versión 7 el software DSpace al tener un modelo de datos cercano al propuesto y un esquema de metadatos flexible que permite extender esquemas preexistentes y la creación de nuevos. Sumado a esto, SEDICI planea la migración a este software en un futuro cercano, lo que facilitaría la eventual integración de la funcionalidad del prototipo en el repositorio. Luego, con modificaciones y extensiones a DSpace, tanto a su *back end* como a su *front end*, se implementaron los servicios propuestos. Los servicios implementados fueron solo una parte del total de servicios propuestos como consecuencia de los tiempos y el marco de la tesina.

Al finalizar el desarrollo, se obtuvo un prototipo funcional con 137 perfiles de autor, cada uno relacionado con las reparticiones UNLP a las que pertenecen y también con sus publicaciones en SEDICI. El prototipo fue desarrollado de manera tal que la integración con SEDICI, una vez que SEDICI haya migrado a DSpace 7, sea sencilla y sin mayores inconvenientes. Asimismo, a partir de la recolección de datos de autores UNLP, se confeccionó una base de datos de 48.890 autores, que permitirá el crecimiento de la cantidad de perfiles de autor a implementar a medida que los datos sean validados.

Capítulo 10 - Trabajos futuros

Si bien se logró el desarrollo de un prototipo de perfil de autor completamente funcional, junto con diversos servicios de valor agregado en torno al perfil, no fue posible implementar todos los servicios propuestos. Y aunque el prototipo fue pensado para su posterior integración en el repositorio SEDICI, todavía no son compatibles dado que el repositorio está desarrollado en una versión anterior de DSpace a la del prototipo. A su vez, el proceso de recolección de datos de los autores UNLP por los distintos repositorios y bases de datos podría volver a utilizarse en un futuro, ya sea con el objetivo de incorporar una mayor cantidad de perfiles a los ya existentes, o incluso para completar e incrementar la base de autoridades de autores de SEDICI; pero para esto ese proceso debería revisarse y ser mejorado e incluso automatizado para poder ser utilizado por una base de datos perteneciente al repositorio institucional. Todos estos puntos y algunos más podrían aplicarse en un futuro, y a continuación se describen todas estas posibles mejoras tomando como base lo desarrollado en la tesina:

Implementación en SEDICI y CIC Digital junto con la integración con la versión final de DSpace 7. Como primer paso, para poner en producción el prototipo para los repositorios de SEDICI y CIC Digital, uno de los objetivos principales de esta tesina, se deben migrar los repositorios a la versión 7 de DSpace, que en la actualidad se encuentran en las versiones 5 y 6 respectivamente. Una vez actualizados los repositorios, se encuentran las condiciones dadas para poner en marcha el prototipo creado; en el caso de SEDICI es más simple porque ya se encuentran realizados todos los relevamientos y recopilación de datos pertenecientes a la Universidad Nacional de La Plata y se puede comenzar con ese *dataset* obtenido para integrar un gran número de autores. En cuanto al repositorio de CIC Digital se debe realizar el proceso completo para obtener el *dataset* de autores de CIC antes de incorporar los perfiles de autores. Se debe tener en cuenta a la hora de realizar el relevamiento de autores que, según la cantidad de centros publicados en el sitio web de la CIC, aproximadamente un

30 % de los centros poseen doble dependencia con la UNLP por lo que ya se parte con un gran *dataset* que contiene una base de autores procesados y duplicados para esta institución.

Implementación de los servicios que quedaron pendientes:

Exposición del perfil en algún formato para interoperar. Desde este punto se pueden evaluar varios formatos para exponer los datos del perfil del autor, como OAI-PMH, OpenSearch o RDF, pero si se piensa en un caso de uso concreto, desde el proyecto de visibilidad web de la Universidad Nacional de La Plata se realizó un desarrollo que permite generar el perfil de autor para los integrantes de las unidades académicas (Villarreal *et al.*, 2017). Para esto se utiliza un *plugin* desarrollado en Wordpress donde se cargan los datos personales y/o profesionales de cada investigador y se recupera a través del protocolo OpenSearch, desde un repositorio compatible con este protocolo, su producción científica. De esta manera se puede pensar en añadir la posibilidad de exponer el perfil en algún protocolo que permita exponer el perfil completo para integrar en los sitios web sin la necesidad de cargar y replicar los datos de cada investigador.

Permitir a un autor autenticarse y administrar tanto sus datos de usuario como parte de su perfil de autor. Queda pendiente la posibilidad de implementar el servicio que permita relacionar al registro *e-person* de DSpace con el perfil del autor para que cada usuario pueda administrar sus datos. En este punto se podrían evaluar algunas soluciones como utilizar Single Sign On con sistema de acceso único de la UNLP o con ORCID.

Permitir a un usuario sugerir cambios o pedir la corrección de un dato erróneo de un perfil de autor que sólo deben ser editados por los administradores del repositorio. Este servicio está pensado para que el autor pueda solicitar correcciones que sólo pueden editar los administradores del repositorio, como puede ser la adjudicación de una publicación ya que es un dato sensible y no se deberían permitir errores. Algo similar a lo que hace Scopus cuando se requieren modificaciones de los perfiles de autores. Se puede plantear como uno de los primeros servicios a implementar, ya que podría servir como una primera etapa para solucionar el punto anterior, en la que un autor pueda enviar una solicitud a

los administradores del repositorio para poder modificar la totalidad de los datos del perfil hasta que efectivamente se implemente la autenticación.

Integrar con otros servicios de perfil de autor e identificadores persistentes para el intercambio de metadatos. La idea de este servicio es conectarte con la API REST de alguno de los principales proveedores para detectar cambios en los datos cargados en sus distintos perfiles y replicar estos cambios en el perfil implementado en el repositorio.

Mejorar el proceso de deduplicación de los datos de los autores. Mientras más precisión se obtenga durante este proceso, menor será el tiempo de revisión manual y de importación. Como ya se vio se han encontrado algunos falsos positivos a la hora de indicar que dos autores son la misma persona y otros falsos negativos, es decir que no se detectó que un autor es el mismo en distintas fuentes. La deduplicación por nombre es un problema cuando no se tienen muchos datos extra que ayuden a identificar a un mismo autor. En un análisis rápido, se puede ver claramente que la mayoría de los autores cuentan con la institución, pero el problema es que, en muchos de los casos, estos datos son cargados por los mismos autores de manera manual, por lo que una misma institución puede aparecer con distintos nombres dependiendo de cómo el autor la haya cargado, por ejemplo, la institución “Universidad Nacional de La Plata” puede aparecer con su nombre completo o bien solo con sus siglas (UNLP). Por este motivo se tomó el *dataset* del módulo de autoridades de SEDICI y se realizó una deduplicación de todas las instituciones con este *dataset*. Si bien, en el proceso durante la tesina se consideró como una buena solución, este punto se puede reforzar para mejorar el proceso de duplicación. Como primera medida se debe mejorar el *dataset* de instituciones. Un buen punto a tener en cuenta es que el número de instituciones pertenecientes a la Universidad Nacional de La Plata es lo suficientemente acotado como para considerar realizar el trabajo de completado de los datos de forma manual, de manera que permita tener una fuente con datos validados. De esta manera se pueden tomar los datos disponibles de las instituciones, completar las faltantes y agregar variaciones de cada una que ayuden a la herramienta Dedupe a encontrar duplicados. Otra solución que puede ayudar a encontrar duplicados desde las instituciones es tener en cuenta el árbol

de jerarquía de las instituciones, se han encontrado casos de falsos negativos donde un mismo autor no fue detectado como duplicado, cuando por ejemplo, en una fuente se encontraba como institución un departamento y en la otra fuente un laboratorio perteneciente a ese mismo departamento.

Incluir el resto de los autores UNLP en el armado de los perfiles. Si bien se tienen todos los autores deduplicados con el proceso mencionado en la tesina, no sería una buena práctica importar todos los autores juntos dentro del repositorio, debido a que esto requiere una revisión manual por parte de los administradores para validar los datos antes de integrarlos. Por este motivo se plantea como trabajo futuro, que luego de migrar SEDICI a DSpace 7 e integrar el servicio de perfiles autores en el repositorio, se diseñe una importación de autores por etapas, donde se puede utilizar un criterio similar al que se utilizó al importar los autores de la Facultad de Ciencias Naturales y Museo dentro del prototipo planteado. También se pueden evaluar distintos objetivos a la hora de plantear las etapas de migración, como generar los perfiles de los autores con una mayor trayectoria, por lo que se deberán importar los perfiles de los autores con más publicaciones dentro del repositorio, o tener los perfiles generados de los autores que se encuentran publicando de manera activa, por lo que se debería importar primero los perfiles de quienes realizaron publicaciones en los últimos años. Este punto se encuentra relacionado con el punto anterior, pues el hecho de mejorar el proceso de deduplicación de datos afecta directamente a la importación: si se mejoran los resultados a la hora de deduplicar los autores, se reduce el tiempo de revisión manual de datos y en consecuencia se reduce el tiempo en el que se logra la importación de los autores al repositorio.

Referencias

Ali, Sabha, Jan, Sumaira & Amin, Iram. (2019). Status of Open Access Repositories: a Global perspective.

Alonso Arévalo, J., Subirats Coll, I. & Martínez Conde, M. (2008). Informe APEI sobre acceso abierto. (Informe APEI; 2). Gijón: Asociación Profesional de Especialistas en Información. http://eprints.rclis.org/12507/1/informeapeiaccesoabierto.pdf

Artigas, W., & Casanova, I. (2020). Influencia de las redes sociales académicas en la construcción de la identidad digital latinoamericana. *Anales de Documentación*, 23(2). https://doi.org/10.6018/analesdoc.397551

Barrueco Cruz, J. M. & Navalón, J. A. (2015). Desarrollo de perfiles de autores en DSpace para el repositorio institucional de la Universitat de València. https://helvia.uco.es/handle/10396/12622/

Confederación de Repositorios de Acceso Abierto (COAR). (2013, October 23). Incentivos, integración y mediación: Prácticas sostenibles para poblar repositorios. COAR. https://www.coar-repositories.org/files/Sustainable-best-practices-spanish_final.pdf

De Giusti, M. R., Villarreal, G. L., Terruzzi, F. A., Oviedo, N. F., & Lira, A. J. (2013, Agosto). Interoperabilidad entre el Repositorio Institucional y servicios en línea en la Universidad Nacional de La Plata. PKP International Scholarly Publishing Conferences, Mexico, Mexico. http://sedici.unlp.edu.ar/handle/10915/27406

District Data Labs. (2018, June 12). Basics of Entity Resolution with Python and Dedupe - District Insights. Medium. https://medium.com/district-data-labs/basics-of-entity-resolution-with-python-and-dedupe-bc87440b64d4

DSpace. (n.d.). Functional Overview - DSpace 6.x Documentation - LYRASIS Wiki. Lyrasis Wiki. https://wiki.lyrasis.org/display/DSDOC6x/Functional+Overview

DSpace 7 - Configurable Entities. (n.d.). Google Docs. https://docs.google.com/document/d/1X0XsppZYOtPtbmq7yXwmu7FbMAfLxxOCONbw0_rl7jY/edit

Duraspace. (2018, May 15). duraspace/pcdm. GitHub. https://github.com/duraspace/pcdm/wiki

Enríquez, J. G., Domínguez-Mayo, F. J., Escalona, M. J., Ross, M., & Staples, G. (2017). Entity reconciliation in big data sources: A systematic mapping study. *Expert Systems with Applications*, 80, 14–27. https://doi.org/10.1016/j.eswa.2017.03.010

EuroCRIS. (n.d.). Main features of CERIF. https://www.eurocris.org/cerif/main-features-cerif

European Commission. (2019). Trends for open access to publications. European Commission Website. https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/trends-open-access-publications_en

Ferreira, Anderson, Gonçalves, Marcos & Laender, Alberto. (2013). Disambiguating Author Names in Large Bibliographic Repositories.

Ferreras Fernández, T. (2013). La interoperabilidad: el pegamento técnico para unir repositorios.

Froufe, N. Q. (2016). La emergencia de las redes sociales académicas: su impacto académico. *Opción*, 32(10), 517-528.

García, N. E., Jaroszczuk, S. E., & de Bibliotecología, C. (2010). Objetos digitales: una experiencia de representación con metadatos Dublin Core. de Encuentro Nacional de Catalogadores (1: 2008: Buenos Aires). I Encuentro Nacional de Catalogadores: experiencias en la organización y tratamiento de la información e, 1, 193-206.

García Gómez, C. (2012). ORCID: un sistema global para la identificación de investigadores. *El Profesional de la Información*, 21(2), 210-212.

García Peñalvo, F. J. (2017). Cómo Mejorar La Visibilidad De La Producción Científica. El Perfil del Investigador. https://repositorio.grial.eu/handle/grial/910

García Peñalvo, F. J. (2018). Identidad digital como investigadores. La evidencia y la transparencia de la producción científica [Digital Identity as Researchers. The Evidence and Transparency of Scientific Production]. *Education in the Knowledge*

Society, 19(2), 7-28. doi:10.14201/eks2018192728doi:10.14201/eks2018192728:10.14201/eks2018192728

Genovés, P. (2017). Perfiles de autor en repositorios institucionales. *Palabra Clave* (La Plata), 7(1), e033. http://sedici.unlp.edu.ar/handle/10915/63375

Keeping up with RIMS. (n.d.). Association of College and Research Libraries. Retrieved August 24, 2020. http://www.ala.org/acrl/publications/keeping_up_with/rims

Krämer, T., Momeni, F., & Mayr, P. (2017). Coverage of author identifiers in Web of Science and Scopus. arXiv preprint arXiv:1703.01319

Latif, Atif, Borst, Timo & Tochtermann, Klaus. (2018). Compiling Scholarly Profile Pages by Integrating External Authority Data. 411-412. doi:10.1145/3197026.3204473.

Lemberget, Tove, Meade, Clair, Morton, Stephanie, Gibbons, Anna, & Ferguson, Kathryn (2017). Keeping an institutional repository relevant. In: Posters from the Open Repositories Conference. 260. From: OR2017: Open Repositories Conference, 27-30 June 2017, Brisbane, QLD, Australia.

Lorenzo Escolar, N. & Pastor Ruiz, F. (2012). Un análisis de los principales sistemas de identificación y perfil para el personal investigador. *Aula Abierta*, 40(2), 107-118. https://dialnet.unirioja.es/servlet/articulo?codigo=3921021

Lyrasis. (n.d.). DSpace-CRIS Home. LYRASIS Wiki. https://wiki.lyrasis.org/display/DSPACECRIS

Meadows, A. (2017, August 25). Ten reasons to get —and use— an ORCID iD. Elsevier. https://www.elsevier.com/connect/authors-update/ten-reasons-to-get-and-use-an-orcid-id

ORCID. (2021, March 19). About. https://info.orcid.org/what-is-orcid/

Oviedo, N. F. (2012, November). Curso avanzado de capacitación en DSpace [Objeto de conferencia]. Curso de capacitación (Alerta al conocimiento), Santiago de Chile, Chile. http://sedici.unlp.edu.ar/handle/10915/25304

Vargas-Arcila, A. M., Baldassarri, S., & Arciniegas, J. L. (2016). Análisis de Esquemas de Metadatos para la Marcación de Contenidos Educativos. *Formación Universitaria*, 9(5), 85-96.

Villarreal, G. L., Manzur, E., Vila, M. M., & De Giusti, M. R. (2017, October). Interoperabilidad con repositorios digitales: uso de OpenSearch en sitios web institucionales. In Conferencia Internacional sobre Bibliotecas y Repositorios Digitales de América

Latina (BIREDIAL-ISTEC'17) y Simposio Internacional de Biblioteca Digitales (SIBD'17)(La Plata, 2017) (Vol. 7). http://sedici.unlp.edu.ar/handle/10915/63566

Anexos

Anexo 1 - Detalle de los 28 repositorios del ranking de Webometrics con perfiles de autor, junto con los servicios que estos ofrecen

Repositorio	Ejemplo	Software utilizado	Publicaciones	Estadísticas	Redes sociales académicas	Datos personales	Afiliación	ORCid o algún identificador persistente
https://europepmc.org/	Link	EMBL-EBI	Si	Si	Si	No	Si	- Orc id (se utiliza también como id interno)
https://ideas.repec.org/	Link	Desarrollo propio	Si	Si	No	Si	Si	No
https://philpapers.org	Link	Centre for Digital Philosophy, University of Western Ontario	Si	No	No	Si	Si	No
https://www.narcis.nl/	Link	Cooperation project of KNAW Research Information	Si	No	No	No	No	- Rearcher id - Scopus id - Orc id
https://air.unimi.it/	Link	Dspace	Si	Si	No	No	Si	No
https://ir.nctu.edu.tw/	Link	Dspace	Si	Si	No	Si	Si	No
https://digital.csic.es/	Link	Dspace Cris	Si	SI	SI	Si	Si	- Orc id - Scopus id - Researcher id
https://riunet.upv.es/	Link	Dspace	Si	No	No	Si	Si	- Orc id - Scopus id - Researcher id

http://s-space.snu.ac.kr	Link	Dspace	Si	No	No	Si	Si	No
http://library.ums.ac.id/	Link	Wordpress	Si	No	No	No	No	No
https://ddd.uab.cat/	Link	Invenio software	Si	Si	No	Si	Si	No
https://uclouvain.be/	Link	Drupal	Si	No	No	Si	Si	No
https://biblio.ugent.be/	Link	Ghent University Library	Si	No	No	Si	Si	No
http://english.opt.cas.cn/	Link	Qingyun Software	No	No	No	Si	No	No
http://repository.vnu.edu.vn/	Link	Dspace	Si	Si	No	Si	Si	No
https://repository.kpfu.ru/eng/	Link	Desarrollo propio	Si	No	Si	Si	Si	- Orc id - Researcher id - Scopus id - Elibrary id
https://koasas.kaist.ac.kr/	Link	Dspace	Si	Si	Si	Si	Si	- Researcher id
https://eprints.qut.edu.au	Link	Eprints	No	No	No	Si	Si	-Orcid
http://uvadoc.uva.es/	Link	Dspace	Si	No	Si	Si	Si	- Scopus id - Orc id
https://digitalcommons.usu.edu/	Link	Digital Commons	Si	No	No	Si	Si	No
https://orbit.dtu.dk/	Link	Pure	Si	Si	No	Si	Si	No
https://ri.conicet.gov.ar/	Link	Dspace	Si	No	No	Si	Si	No
https://ir.uiowa.edu/	Link	Digital Commons	Si	No	No	Si	Si	No

https://archive-ouverte.unige.ch/	Link	Desarrollo propio	Si	Si	No	No	No	No
https://scholarworks.umass.edu/	Link	Digital Commons	Si	No	No	Si	Si	No
https://repositorio.uam.es/	Link	Dspace	No	Si	Si	No	Si	- Orcid - Researcher id - Scopus id
https://repository.upenn.edu/	Link	Digital Commons	Si	No	No	Si	Si	No
http://koara.lib.keio.ac.jp/	Link	Keio Media	Si	No	Si	Si	Si	No

Anexo 2 - Listado de las bases de datos desde donde se extrajo información de los autores UNLP

Nombre y URL	Información obtenida	Autores encontrados	Técnica de recopilación de los datos	¿Se incluyó en el armado de la base de datos final?	Comentarios
Base de autoridades SEDICI http://sedici.unlp.edu.ar/	Nombre, apellido, ID interno del autor en SEDICI, e-mail, DNI, CUIT, ORCID, enlace a los perfiles de Google Scholar y ResearchGate, filiación, ID de la institución en SEDICI y comentarios sobre el autor	24276	Se solicitó una copia de la base de datos	Sí	Base de datos de las autoridades del repositorio SEDICI. Enorme cantidad de autores con variados datos pero muchos de ellos incompletos. No todos son de la UNLP, pero a partir de la filiación se puede deducir. Algunos datos pueden estar desactualizados, por lo general pasa cuando los autores cambian de dependencia o se convierten en una doble dependencia.

Usuarios SEDICI http://sedici.unlp.edu.ar/	Sólo e-mail		Se solicitó una copia de la base de datos	No	Base de datos con los usuarios de SEDICI, sólo contiene el e-mail (ni siquiera el nombre), por lo que se descartó incluirla a la hora de procesar y cruzar datos para encontrar duplicados.
Base de autoridades CIC Digital https://digital.cic.gba.gob.ar/	Nombre, apellido, filiación CIC, CUIT, DNI, categoría de investigador CIC, ID interno del autor en CIC Digital		Se solicitó una copia de la base de datos	No	Tipo y calidad de los datos similares a los de la base de autoridades de SEDICI, pero muchos de los autores eran solo de la CIC y no de la UNLP. Además, las bases de autoridades son gestionadas por los mismos equipos, por lo que los autores cargados en CIC pertenecientes a la UNLP ya se encuentran en la base de autoridades de SEDICI.
CONICET Digital https://ri.conicet.gov.ar/	Nombre, apellido y filiación		Cosecha OAI-PMH	No	Se obtuvieron los datos de los autores desde una cosecha OAI-PMH. Los datos recuperados no fueron un gran aporte a la hora de encontrar duplicados en otras fuentes. Sólo se encontraba el nombre y la filiación, que en la mayoría la indicada era solo la UNLP, sin incluir instituciones hijas, lo que reduce las probabilidades de encontrar un registro duplicado.
Google Scholar https://scholar.google.com/	Nombre, apellido, filiación, dominio del e-mail del autor, cantidad de citas en Google Scholar, ID y URL del autor en Google Scholar	4496	Web scraping	Sí	Base de datos con los autores UNLP en Google Scholar. Estos datos son altamente confiables ya que se tiene en cuenta el dominio del e-mail del autor para saber si es o no UNLP. Además, el mismo autor es el que ingresa los datos de su filiación. La URL de Google Scholar sirve como identificador unívoco.
SIGEVA UNLP https://sigeva.unlp.edu.ar			Se solicitó una copia de la base de datos	No	No se obtuvieron los datos solicitados.
SIGEVA CIC https://cic.sigeva.gob.ar			Se solicitó una copia de la base de datos	No	No se obtuvieron los datos solicitados.

SIBIPA http://sibipa.cic.gba.gob.ar			Se solicitó una copia de la base de datos	No	No se obtuvieron los datos solicitados.
Nómina del personal UNLP https://unlp.edu.ar/			Se obtuvo desde la página web en formato PDF	No	No se pudieron obtener los datos en un formato amigable. De todas maneras esta base de datos sólo tenía el nombre y apellido de los autores por lo que no generaba un gran aporte a la hora de detectar registros duplicados y agregar información extra para los autores.
Participantes en Proyectos de extensión http://proyectos-extension.unlp.edu.ar/	Nombre, apellido, e-mail	1413	Se obtuvo desde la página web	Sí	Base de datos con pocos datos sobre los autores pero muy confiable al provenir directamente de la UNLP. No todos los autores, sin embargo, pertenecen a la UNLP. Provee un identificador unívoco como es el e-mail.
Participantes en Proyectos de extensión 2 http://proyectos-extension.unlp.edu.ar/	Nombre, apellido, -mail, DNI, filiación, teléfono, rol en filiación	14038	Web scraping	Sí	Base de datos más completa que la obtenida mediante la funcionalidad de descarga de los datos desde la página web.
Integrantes de Proyectos SECYT 2013	Nombre, apellido, facultad, e-mail	3791	Se solicitó una copia de la base de datos	Sí	Base de datos con pocos datos sobre los autores (a diferencia de proyectos de extensión, se tiene también la facultad del autor) pero muy confiable al provenir directamente de la UNLP. Uno de los datos es el e-mail, que puede ser usado como identificador unívoco.
ResearchGate https://www.researchgate.net	Nombre, apellido, título profesional, biografía, departamento al que pertenece el autor dentro de la UNLP, cargo, fechas inicio y fin del cargo, lenguajes que habla, URL	3340	A través de la API REST de ResearchGate, junto con un script Python	Sí	ResearchGate tiene una gran cantidad de datos y mucha información sobre los autores, sin embargo no tiene ningún identificador unívoco. Los datos son relativamente confiables porque el autor mismo es el que los carga. Como dato interesante tiene la fecha inicio y fecha de fin en la que el autor tuvo una determinada filiación.

	de la foto del autor y áreas de interés				
Scopus https://www.scopus.com/freelookup/form/author.uri	Scopus ID, ORCID, EID, área de investigación, nombre, apellido, filiación, variantes del nombre, cantidad de documentos en Scopus	3399	A través de la API REST de Scopus, junto con un script Python	Sí	Scopus provee de varios datos interesantes y una gran cantidad de autores. Además del Scopus ID, también cuenta con el ORCID de algunos autores. La confiabilidad de los datos, como ResearchGate, es bastante alta dado que es uno de los identificadores persistentes a nivel mundial.
Lens.org https://www.lens.org/				No	No se encontró información propia de los autores, más que su nombre y apellido.
Dimensions https://dimensions.freshdesk.com				No	No se pudo acceder a la información de los autores, si es que estos tienen información propia en este portal.
Academia.edu https://unlp.academia.edu/	El único dato que provee es el departamento, además del nombre y apellido, y no es confiable. Por ejemplo tiene como nombres de departamento a "Art", "Arte", "Arte". Solamente se encontró una sola persona con e-mail y biografía			No	Poca cantidad de datos de los autores y solo posee el nombre del departamento del autor, además de su nombre. Los datos son poco confiables, se encontraron departamentos mal escritos o erróneos. También tiene como dato en algunos casos la biografía del autor, pero estos casos son extremadamente pocos. No se incluyó esta base de datos en el conjunto final para analizar, debido a su poca confiabilidad y poco aporte.
Acta Académica https://www.aacademica.org/unlp.principal/tabs/profiles	Nombre, apellido, filiación, biografía y foto	7		No	Aunque con algún que otro dato interesante, la cantidad de autores que se recuperaron era ínfima en comparación con las demás bases de datos a analizar.

ORCID https://orcid.org	ORCID, nombre biografía, URL a páginas web del autor, e-mails, palabras clave sobre el autor, otros identificadores (por ejemplo, SCOPUS ID), información de la historia académica del autor, filiaciones junto con el cargo que ocupó en cada filiación, con fecha de inicio y fecha de fin	2562	A través de la API REST de ORCID, junto con un script Python	Sí	Desde ORCID se obtuvieron en algunos casos, además del propio ORCID, otros identificadores persistentes o unívocos del autor, como el e-mail, el SCOPUS ID, la URL de Google Scholar o el Researcher ID. Esto, la gran cantidad de autores obtenidos y el resto de la información que ofrece la hace una base de datos valiosa. Además, al ser el propio autor el que carga los datos, como en ResearchGate, estos son bastante confiables.
PubMed https://pubmed.ncbi.nlm.nih.gov/				No	No se pudo extraer información de los autores. No tiene búsqueda por autor ni información sobre estos, además de su nombre.
arXiv https://arxiv.org/				No	No se pudo extraer información de los autores. No tiene búsqueda por autor ni información sobre estos además de su nombre.
Unpaywall http://unpaywall.org/welcome				No	No fue posible extraer información sobre los autores desde esta fuente.
Publons (Researcher ID) https://publons.com/ https://www.researcherid.com/	Nombre del autor con sus variantes, Researcher ID, ORCID, filiaciones, biografía, URL a una foto del autor, truID, URL a las páginas web de las filiaciones	137	Web scraping	Sí	Como en ORCID y ResearchGate los datos son cargados por el mismo autor, lo que hace que sea una fuente relativamente confiable. No tiene una gran cantidad de autores, pero sí algunos identificadores persistentes..
RePEc Authors https://ideas.repec.org/i/e.html	Nombre, e-mail, página web del autor, dirección, teléfono del autor y filiación junto con su	52	Web scraping	Sí	Poca cantidad de autores pero con datos interesantes. Mucha información personal, incluyendo el e-mail. No se obtuvo información que permita asegurar la confiabilidad de los datos.

	página web, e-mail, teléfono y dirección				
HAL https://hal.archives-ouvertes.fr				No	No tiene información útil de los autores.
Crossref https://www.crossref.org/				No	No se pudo extraer información útil de los autores.
Open Access Button https://openaccessbutton.org/				No	No tiene información de los autores.
Microsoft Academic https://academic.microsoft.com				No	Tiene información de autores UNLP pero no se lograron extraer los datos. Desde este enlace https://academic.microsoft.com/institution/874386039/authors?pi=1, se pueden ver los autores más relevantes pero no acceder a sus datos.
Astrophysics Data System (AdsAbs) https://ui.adsabs.harvard.edu/				No	No se pudo extraer los datos de los autores y tampoco tiene datos además del nombre y apellido.