Tesina de licenciatura

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

Ezequiel Manzur
Santiago Tettamanti

Directora
Marisa R. De Giusti

Asistentes Profesionales
Lic. Ariel Jorge Lira y Dr. Gonzalo Luján Villarreal

UNIVERSIDAD NACIONAL DE LA PLATA

SEDICI
sedici.unlp.edu.ar

Índice

¿Qué se hizo?

1 **Servicio de perfiles de autor** para repositorios institucionales, que incluye datos de autores de la **UNLP.**

Se creó un prototipo funcional con **137** perfiles de autor sobre **DSpace 7**

2 Se **relevaron** los primeros **200 repositorios** del ranking web de transparencia de Webometrics

3 **Se recolectaron datos de autores UNLP**

Se confeccionó una base de datos de **48.890 autores**
Se deduplicaron más de **8000 autores**

Motivación

- El uso de los **repositorios** es una herramienta clave para la difusión de la producción científica
- La mayoría de los repositorios carece de un espacio que centralice y exponga la producción académica de sus autores junto a su información personal/profesional.
- Tener este espacio permite...
    - Exponer producción científica y datos de los autores
    - Servicios de valor agregado
    - Ampliación el impacto de su producción académica
    - Incentivo para que los autores depositen sus trabajos en acceso abierto
    - Mejora de su identidad digital
- **SEDICI** contiene la producción científica de gran parte de los investigadores de la UNLP

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

- Relevar y analizar servicios existentes con respecto a la gestión de autores y sus datos en repositorios institucionales.
- Diseñar una solución apropiada para el repositorio SEDICI, e implementarla en la herramienta DSpace.
- Incorporar al repositorio mencionado el servicio de perfiles públicos de autores

Objetivos

Objetivos específicos

- Recuperar, normalizar y combinar los conjuntos de datos de autores a fin de construir una base de autores de nuestra Universidad.
- Implementar servicios de valor agregado en los perfiles de autor.
- Describir una solución general aplicable en otros repositorios institucionales.

Contexto

Contexto

Repositorio institucional

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.

SEDICI

Es el **Repositorio Institucional de la Universidad Nacional de La Plata**

- Tiene más de **110000 recursos***
- Desarrollado sobre **DSpace 5**
- Se planea migrar a **DSpace 7** en el corto plazo
- Posee una base de autoridades con más de **25000 autores*.**

Contexto

Perfil de autor

Reúne y **expone** la producción científica y la información personal/profesional de un investigador

- **Enriquecen** la información de un investigador
- Datos personales y de filiación
- **Estadísticas**
- Enlaces a otros identificadores y perfiles de los autores.
- Posibilidad de tener **servicios de valor agregado**

Relevamientos

De repositorios

Análisis de los servicios para autores

De otros sistemas de perfiles de autor

Análisis de los modelos de datos

Relevamientos

De repositorios

Análisis de los servicios para autores

De otros sistemas de perfiles de autor

Análisis de los modelos de datos

¿Dónde se buscó?

Los primeros **200** repositorios del ranking web de transparencia de *Webometrics*.

Objetivo

Conocer el estado general del servicio de perfiles de autor en los repositorios institucionales

¿Qué se relevó?

Cantidad de repositorios con servicio de perfiles de autor, software utilizado, datos mostrados, servicios provistos

Relevamientos

De repositorios

Análisis de los servicios para autores

De otros sistemas de perfiles de autor

Análisis de los modelos de datos

Resultados

- De los **200** repositorios relevados, un **14 %** implementa el servicio de perfiles de autor
- La cantidad de repositorios con perfiles en el ranking creció en el tiempo:

  De 2017 a 2020 aumentó del **10%** al **14%**

Relevamientos

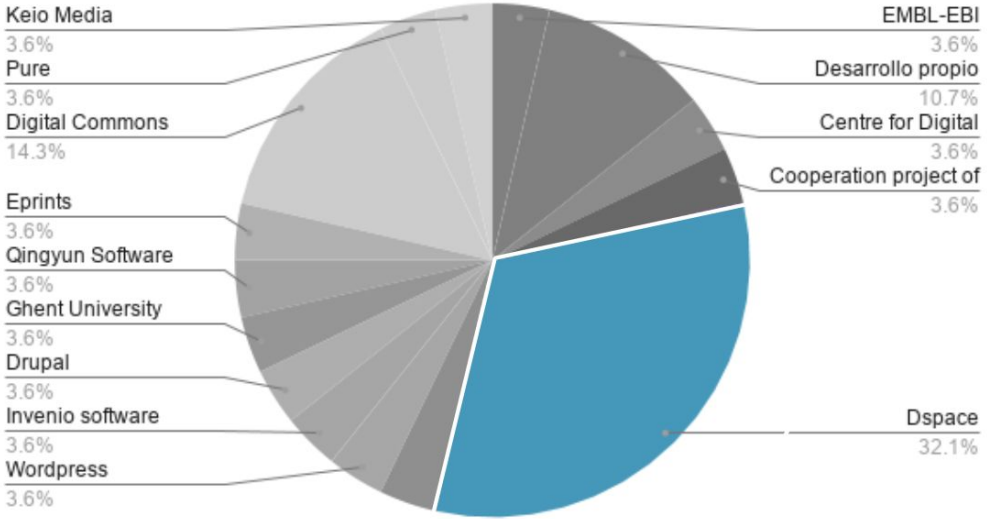
De repositorios

Análisis de los servicios para autores

De otros sistemas de perfiles de autor

Análisis de los modelos de datos

¿Que software usan?



Relevamientos

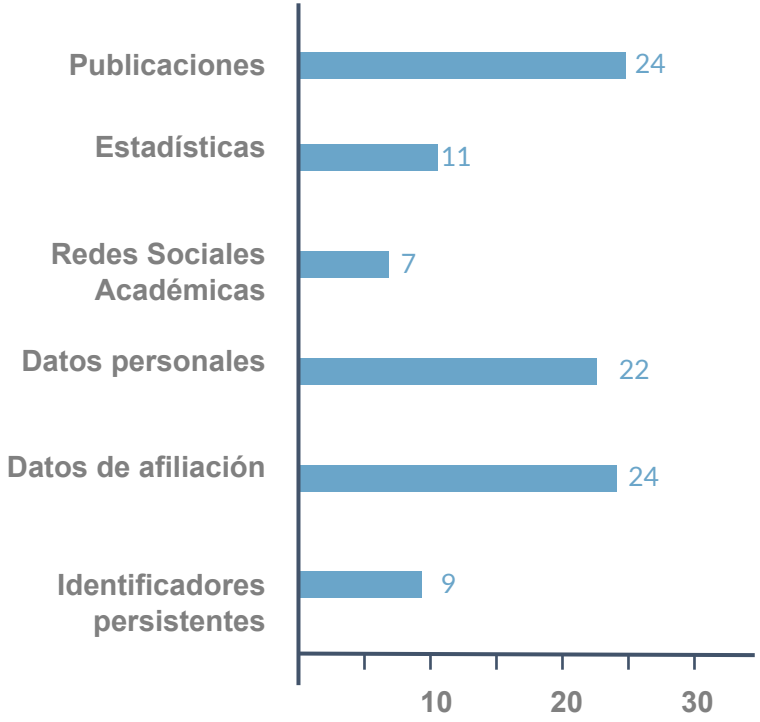
De repositorios

Análisis de los servicios para autores

De otros sistemas de perfiles de autor

Análisis de los modelos de datos

¿Qué servicios implementan?



Relevamientos

De repositorios

Análisis de los servicios para autores

De otros sistemas de perfiles de autor

Análisis de los modelos de datos

Modelos de datos de algunas implementaciones

ORCID

Scopus®

DSPACE

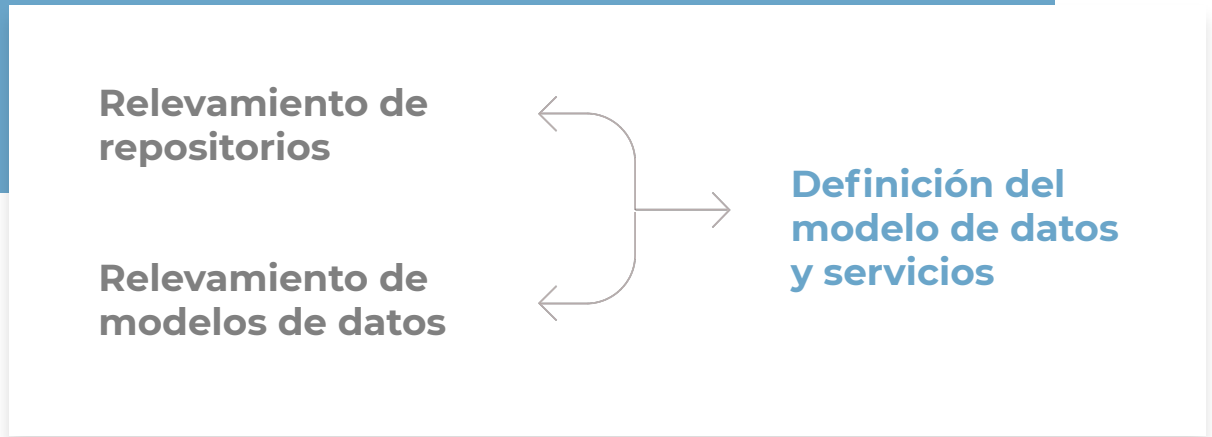
CERIF euroCRIS

DSPACE CRIS

DSPACE 7

Definición de **modelos y servicios**

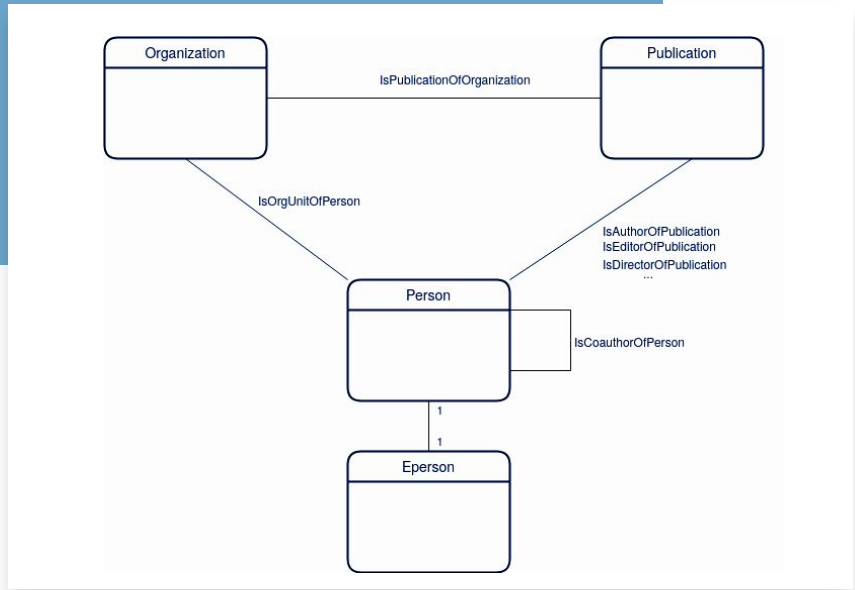
- Ventajas y desventajas de los servicios y modelos relevados
- Aproximación al modelo ideal que se desea implementar.
- En línea con las tendencias observadas



Definición del **modelo**

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

Basado en el modelo de DSpace 7



Definición de los **servicios**

Servicios propuestos para implementar en el prototipo

- Visualización pública del perfil de autor
- Navegación entre las relaciones de un autor, con filtros y búsquedas
- Permitir a un autor autenticarse y administrar tanto sus datos de usuario como parte de su perfil de autor
- Reporte o visualización de estadística

Definición de los **servicios**

Servicios propuestos para implementar en el prototipo

- Exportación del perfil de autor en diversos formatos
- Exponer el perfil en algún formato para interoperar
- Código QR con URL al perfil
- Búsqueda de autores a partir de algún identificador persistente

Metadatos

- Se emplean para describir el **contenido** y otras características de los recursos digitales
- Necesarios para la **búsqueda, gestión** o **recuperación** de los datos

Precisamos metadatos para **personas** y **organizaciones**:

- **DSpace 7** implementa estas dos entidades y extiende los esquemas *Organization* y Person de Schema.org.
- Para el modelo definido se extiende esta base y se definen algunos metadatos específicos, por ejemplo **biografía**, **email** y **variantes en el nombre**.

¿Y los **datos**?

Para poder implementar los perfiles de autor...

Se necesitaban datos de autores **UNLP**

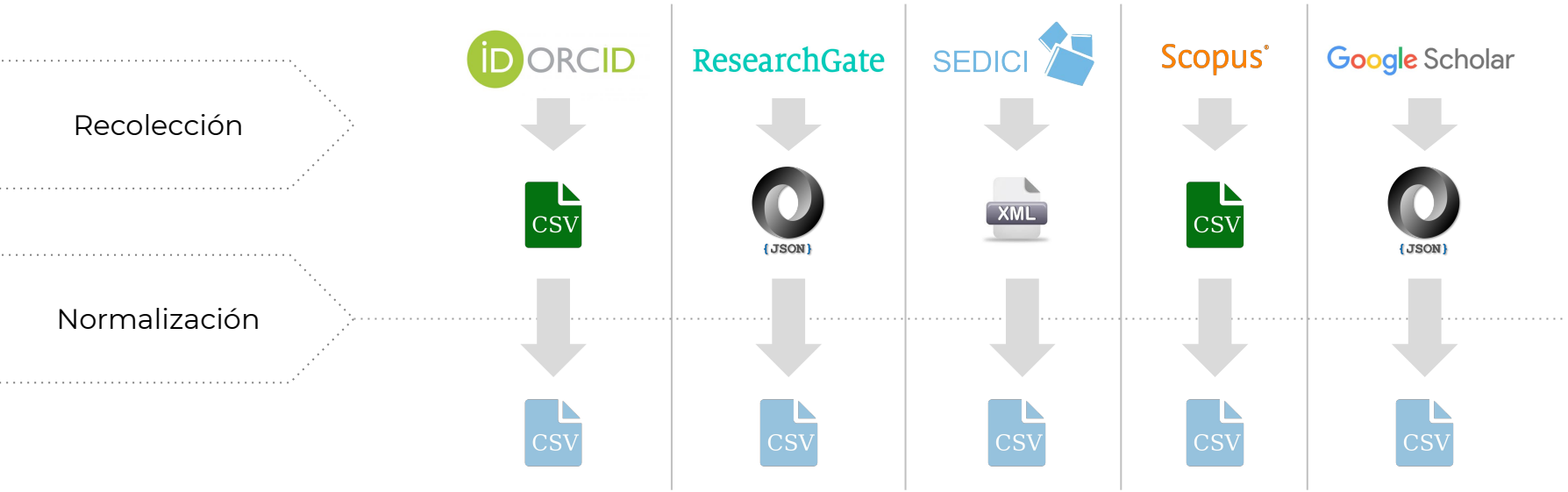
Base de autoridades de **SEDICI**

- Más de **25000** autores
- Datos incompletos
- Algunos duplicados

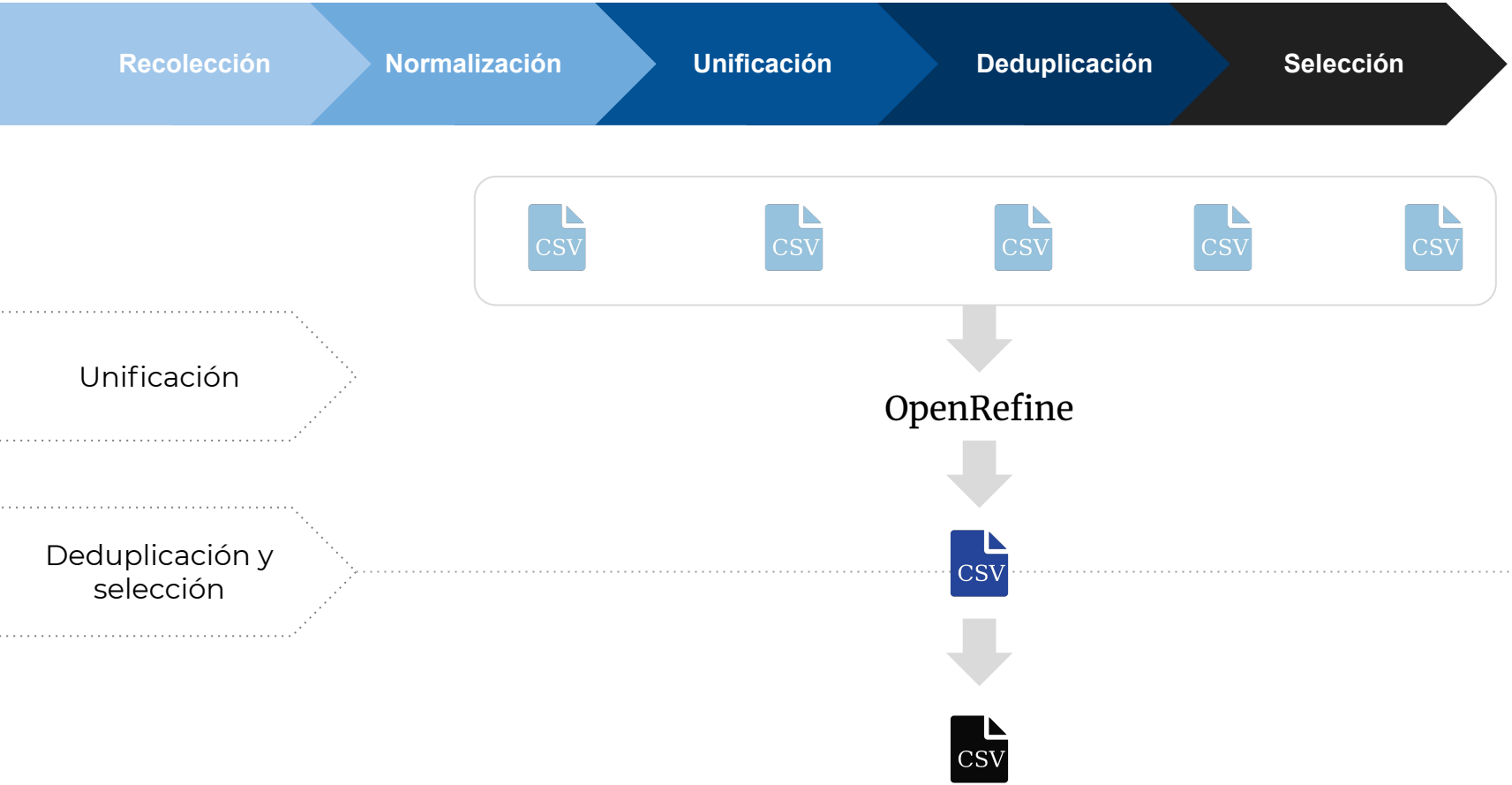
Recopilar datos de autores desde distintas fuentes, **unificarlos, normalizarlos** y **deduplicarlos**

Recolección, normalización y deduplicación de datos

Recopilación normalización y deduplicación de datos



Recopilación normalización y deduplicación de datos



Recolección

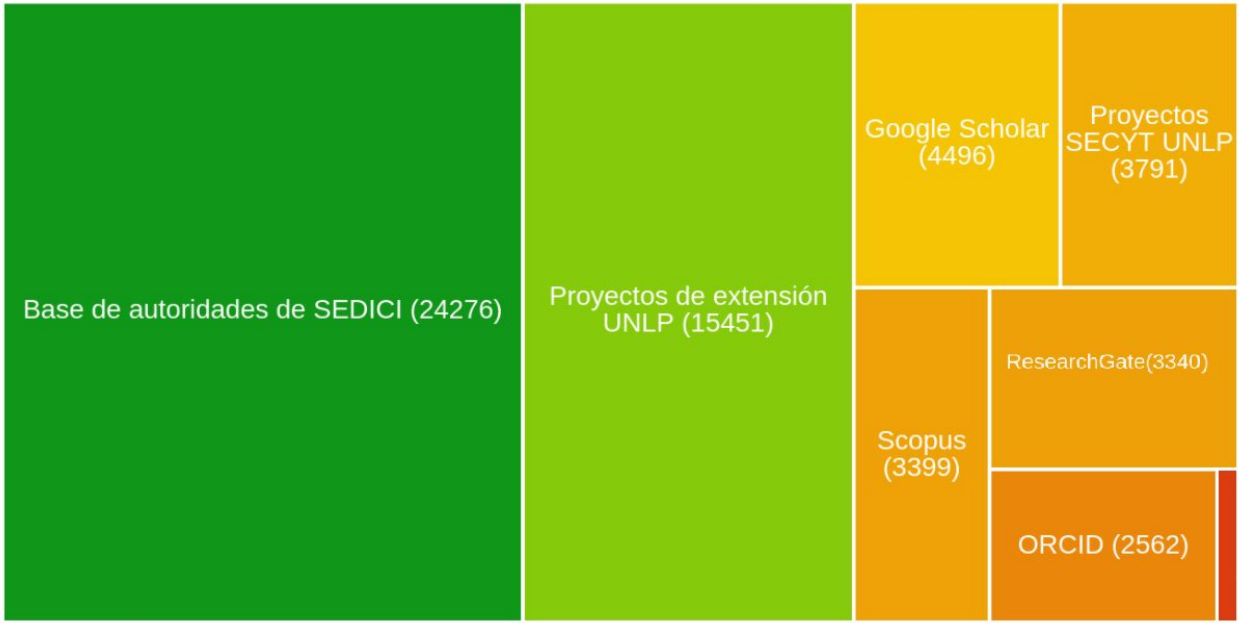
Se buscaron distintas fuentes de datos con información de autores UNLP

- **29** fuentes de datos distintas, solo se obtuvieron datos de 9
  - Base de autoridades **SEDICI, GS, ORCID**, algunas bases de datos de la **UNLP**, entre otros.
- Métodos de recopilación de datos:
  - API REST
  - Web scrapping
  - Solicitud de los datos

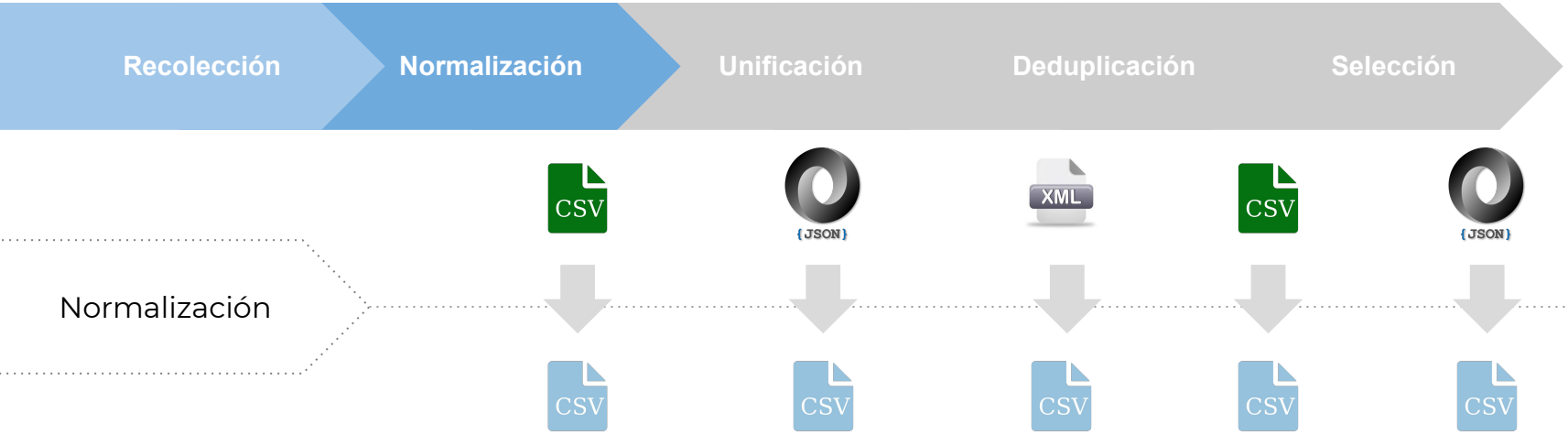
Recolección

Recolección | Normalización | Unificación | Deduplicación | Selección

Autores/ cantidad de autores



Normalización



Luego de recolectar

- Muchos datos en distintos formatos, tanto en la forma de almacenarlos como en el contenido
- Necesidad de unificar los formatos
  - Un único formato (CSV)
  - Normalizar los nombres de las columnas o campos
  - Normalizar contenido (ej celdas multivaluadas)
- Se utilizó **Open Refine**

Unificación

Recolección | Normalización | **Unificación** | Deduplicación | Selección

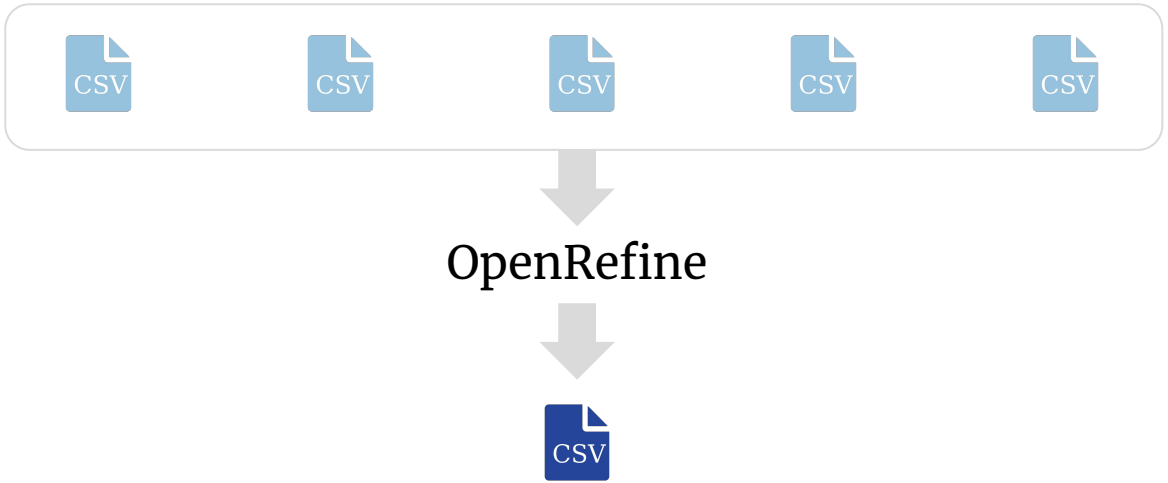
Se unió el contenido de los archivos normalizados en un único archivo

- Se preservó la fuente original
- Se utilizó Open Refine

El resultado

Archivo **CSV**, con **61** columnas (campos) y **57504** filas

Autores repetidos entre distintas fuentes



Deduplicación

Problema: autores repetidos.

¿Cómo resolvemos esto?

- Simple si hay id unívoco (ej orcid)
- Sin id unívoco, utilizando el resto de los datos y complementando con algunas herramientas (ej afiliación y dedupe)

ORCID	APELLIDO	NOMBRE	INSTITUCIÓN	FILE
0000002	Perez	Luciano Ariel	Depto. de Física	orcid_file.csv
0000001	Gonzales	J	Depto. de Economía	orcid_file.csv
	Perez	Luciano	Depto. de Física	sedici file.csv
0000001	Gonzales	Jorge	UNLP	sedici_file.csv

Deduplicación - ID unívoco

Recolección | Normalización | Unificación | Deduplicación | Selección

- Identificar ids unívocos

Por cada tipo de identificador:

- **Ordenar por id**
- Unificar las filas con el mismo id
- Resolver el caso de datos multivaluados

Datos repetidos

Orcid	Apellido	Nombre	Institución	File
0000001	Gonzales	Jorge	UNLP	sedici_file.csv
0000001	Gonzales	J	Depto. de Economía	orcid_file.csv
0000003	Garcia	Maria Laura	Depto. de Física	sedici_file.csv
0000003	García	Laura	Depto. de Física	orcid_file.csv

Deduplicación - ID unívoco

Recolección | Normalización | Unificación | **Deduplicación** | Selección

- Identificar ids unívocos

Por cada tipo de identificador:

- Ordenar por id
- **Unificar las filas con el mismo id**
- Resolver el caso de datos multivaluados

Datos deduplicados

ORCID	APELLIDO	NOMBRE	INSTITUCIÓN	FILE
0000001#0000001	Gonzales#Gonzales	Jorge#J	UNLP#Depto. de Economía	sedici_file.csv#orcid_file.csv
0000003#0000003	Garcia#García	Maria Laura#Laura	Depto. de Física#Depto. de Física	sedici_file.csv#orcid_file.csv

Deduplicación - ID unívoco

Recolección | Normalización | Unificación | **Deduplicación** | Selección

- Identificar ids unívocos

Por cada tipo de identificador:

- Ordenar por id
- Unificar las filas con el mismo id
- **Resolver el caso de datos multivaluados**

Resultados de **57504** a **52878** filas

Datos deduplicados

ORCID	APELLIDO	NOMBRE	INSTITUCIÓN	FILE
0000001	Gonzales	Jorge#J	UNLP#Depto. de Economía	sedici_file.csv#orcid_file.csv
0000003	Garcia#García	Maria Laura#Laura	Depto. de Física	sedici_file.csv#orcid_file.csv

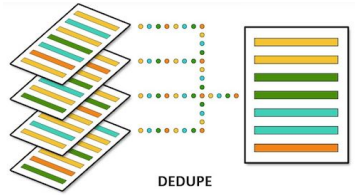
Deduplicación - Nombre, filiación y otros datos

Recolección | Normalización | Unificación | Deduplicación | Selección

Casos en los que los datos son ambiguos

- Ya no hay un dato que identifica unívocamente al autor
- Es el mismo autor, son dos o son tres distintos?

Uso de herramientas **Dedupe** + **Open Refine**



Apellido	Nombre	Institución
Gonzalez	Juan P.	Facultad de Informática#LIFIA
Gonzalez	Juan Pablo	UNLP#LIFIA
Gonzalez	Juan Pablo	Depto. de Física

Deduplicación - Nombre, filiación y otros datos

- Selección de datos a utilizar
  a) Nombre, parte del email y filiación
- Normalización de los datos a utilizar
  a) Una misma filiación aparece con distintos nombres → deduplicación con dedupe
  b) Lista de ocurrencias de nombres para una institución

Apellido	Nombre	Institución
Gonzalez	Juan P.	Facultad de Informática#LIFIA
Gonzalez	Juan Pablo	UNLP#Laboratorio de Investigación y Formación en Informática Avanzada (LIFIA)
Gonzalez	Juan Pablo	Depto. de Física

Deduplicación - Nombre, filiación y otros datos

Recolección | Normalización | Unificación | Deduplicación | Selección

- Con dedupe marcar duplicados
  - Probar varias configuraciones hasta encontrar la adecuada
  - **Se genera un id y valor de confianza por cada autor distinto**
- Unificar las filas repetidas con Open Refine
  - Seleccionar un valor de confianza aceptable
  - Unificar filas con mismo id + valor de confianza aceptable

Resultados

De **52878** a **48890** filas

Apellido	Nombre	Institución	ID	Confianza
Gonzalez	Juan P.	Facultad de Informática#LIFIA	1	0.9
Gonzalez	Juan Pablo	UNLP#LIFIA	1	0.9
Gonzalez	Juan Pablo	Depto. de Física	1	0.4

Deduplicación - Nombre, filiación y otros datos

Recolección | Normalización | Unificación | Deduplicación | Selección

- Con dedupe marcar duplicados
  - Probar varias configuraciones hasta encontrar la adecuada
  - Se genera un id y valor de confianza por cada autor distinto
- Unificar las filas repetidas con Open Refine
  - **Seleccionar un valor de confianza aceptable**
  - Unificar filas con mismo id + valor de confianza aceptable

Resultados

De **52878** a **48890** filas

Apellido	Nombre	Institución	ID	Confianza
Gonzalez	Juan P.	Facultad de Informática#LIFIA	1	0.9
Gonzalez	Juan Pablo	UNLP#LIFIA	1	0.9
Gonzalez	Juan Pablo	Depto. de Física	1	0.4

Deduplicación - Correcciones manuales

Falsos negativos

Falsos positivos

Difíciles de detectar

Se buscaron patrones y causas que permitan encontrar estos errores automáticamente

Errores humanos en el origen de los datos.

- Autores con más de un identificador unívoco.
- Distintos autores con el mismo identificador unívoco.

- Falsos negativos por falta de datos
- Falsos positivos por la herramienta dedupe

Margen de error debido al intervalo de confianza en dedupe.

Selección

¿A qué autores se les genera un perfil?

Se analizaron distintos criterios

- Cantidad de publicaciones en SEDICI **mayor a 20**.
  - Autores de mayor relevancia debían ser los primeros con un perfil (recompensa).
  - Usar al perfil de autor como un **incentivo** para el depósito de obras
  - 1703 autores
- Pertenecientes a la Facultad de Ciencias Naturales y Museo
  - Quedaron **137 autores**
  - Aceptable para su revisión manual en el corto plazo
  - Misma facultad, mayor la posibilidad de relaciones de autoría, bueno para mostrar estadísticas

Selección

Etapa	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

Implementación del prototipo

Basado en **Dspace7 - beta4**

- backend Java + Spring
  - sigue los principios HATEOAS para api REST
- frontend Angular

¿Por qué DSpace (y por qué 7)?

- DSpace es el software que utiliza **SEDICI**

  Se va a migrar a la versión 7 pronto

- El **modelo de datos** concuerda con el modelo de datos propuesto y es flexible para adaptarse.

  Autores como entidades con su propia página a modo de perfil (versiones anteriores no)

- Gran comunidad que lo sustenta.

Implementación del prototipo

Importación de datos

¿Cómo se integran los datos de los autores?

- Migración de la base de datos de SEDICI **reducida** a Dspace7
  - Comunidad Facultad de Ciencias Naturales y Museo
- Extensión del modelo de metadatos
  - Desde la interfaz, para personas y organizaciones
- Creación de las colecciones **destino** en el prototipo
- Importación de los 137 autores

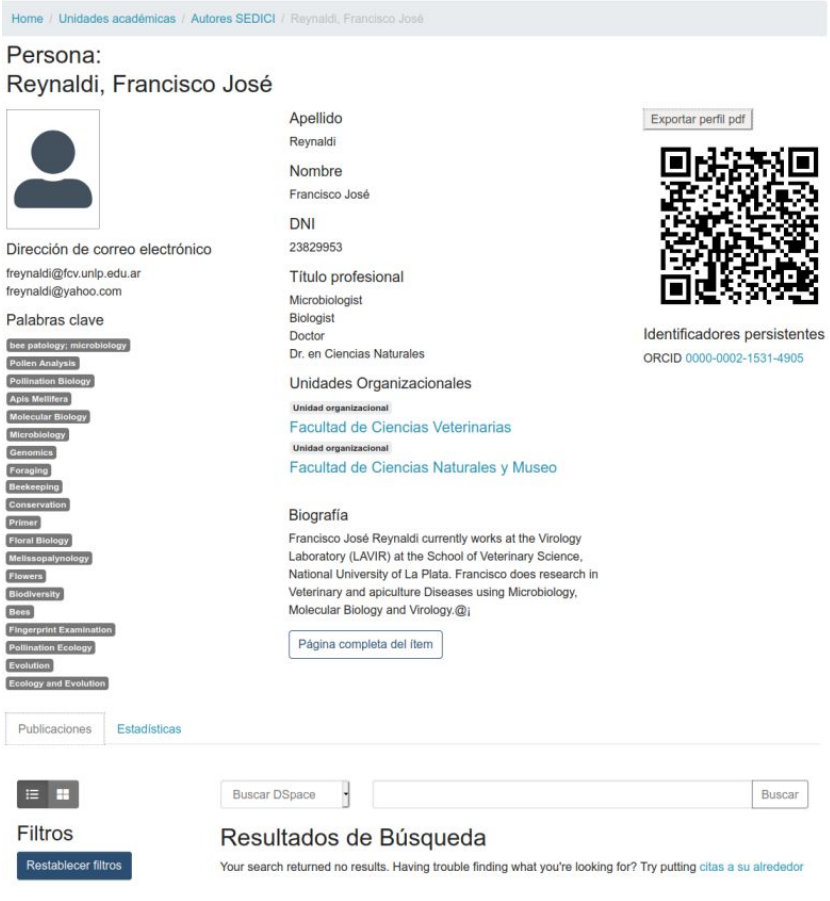
collection	relationship.type	person.givenName	person.familyName
123456789/506	Person	Francisco Raúl	Carnese
123456789/506	Person	Alicia Bibiana	Orden
123456789/506	Person	Gustavo Alberto	Darrigran

- Creación e importación del *.csv* de **instituciones UNLP**
- Creación de las **relaciones** entre las entidades

Implementación del prototipo

Servicios implementados

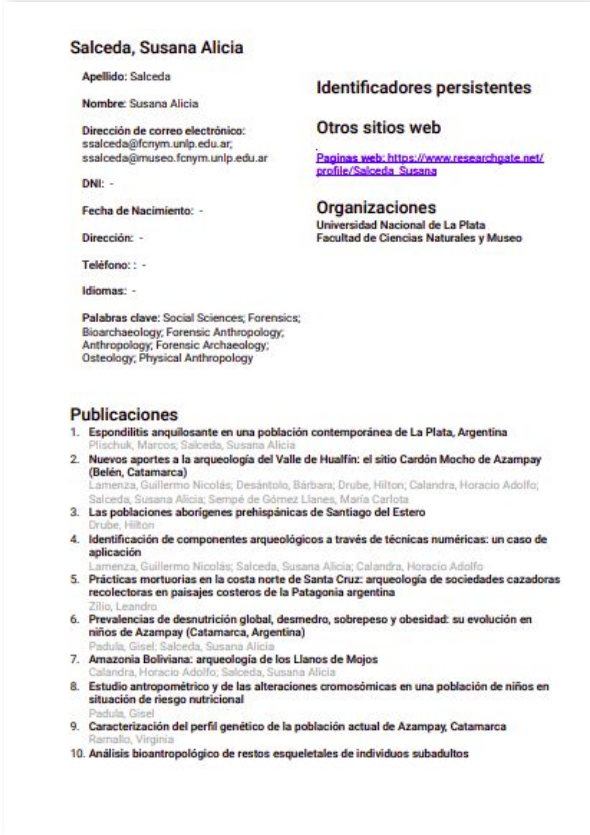
- Visualización del perfil y navegación entre las relaciones del autor
- Visualización de las publicaciones
- Código QR con enlace al perfil de autor



Implementación del prototipo

Servicios implementados

- Exportar perfil en **.pdf**
- Búsqueda en el repositorio a partir de los datos 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

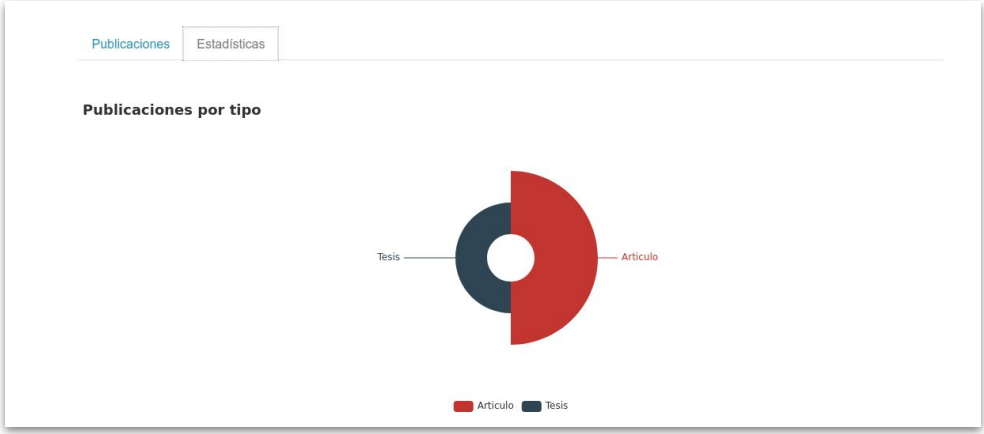
Universidad Nacional de La Plata
Facultad de Ciencias Naturales y Museo

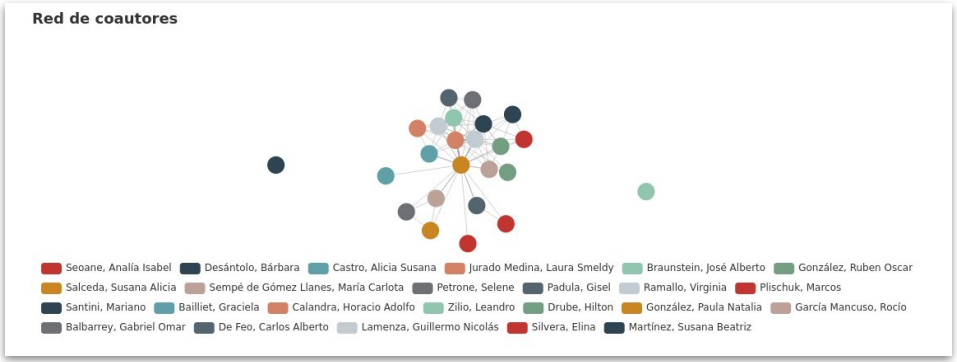
Publicaciones

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**

Implementación del prototipo

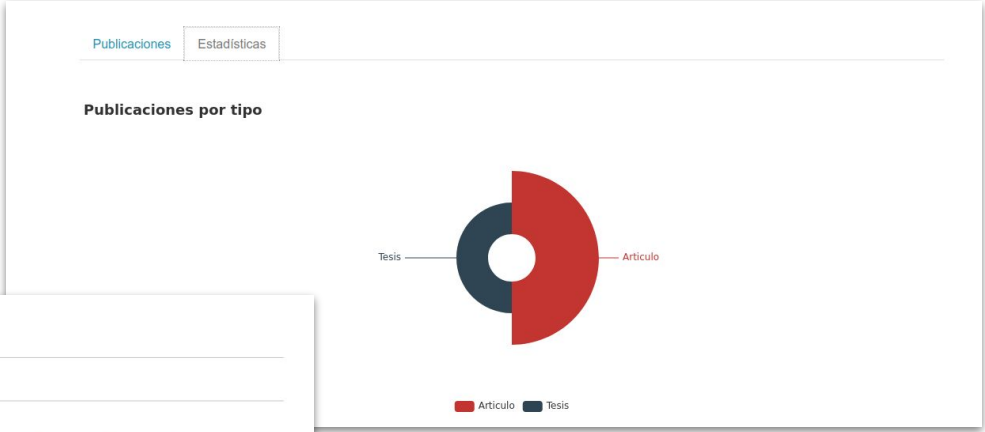
- Visualización de estadísticas

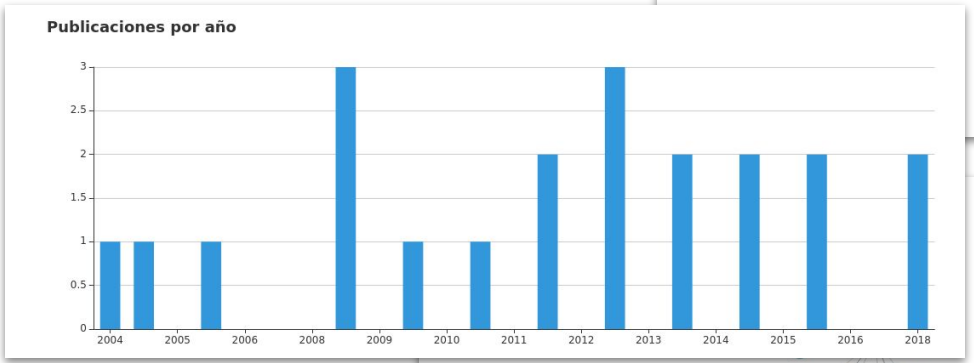


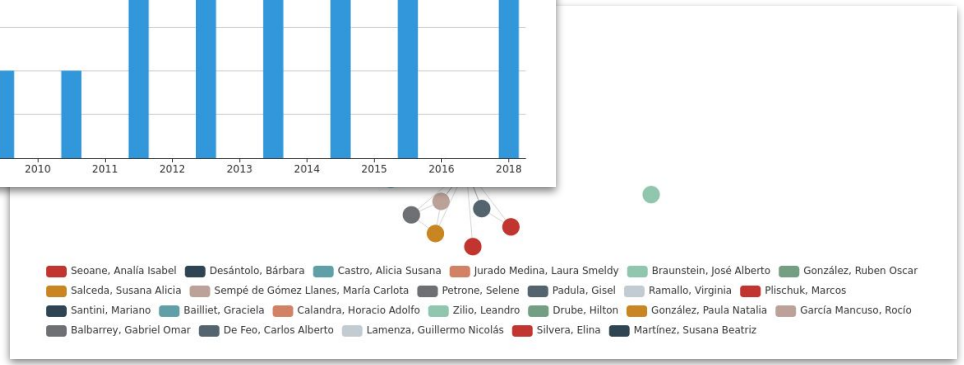


Implementación del prototipo

- Visualización de estadísticas







Conclusiones

- La integración del servicio de perfiles de autor en **SEDICI**:
    - aumentaría la **visibilidad** de los autores y el repositorio
    - Permitiría desarrollo de más **servicios de valor agregado** para los autores
    - Serviría como **incentivo** para que los autores **depositen** más obras en el repositorio
- Al extender de **DSpace 7** la integración del prototipo con cualquier repositorio basado en DSpace 7 debería ser **sencilla** y sin mayores inconvenientes.
- Esa gran cantidad de autores (**48.890**) son una fuente que servirá para:
    - La futura implementación de la **totalidad** **de los perfiles de autor** en SEDICI
    - Completar y aportar mayor cantidad de información a la **base de autoridades** de dicho repositorio

Trabajos futuros

- Implementación en **SEDICI** y **CIC Digital** junto con la integración con la versión final de **DSpace 7**
- Implementación de los servicios que quedaron **pendientes**:
    - Exposición del perfil en algún formato para interoperar
    - Permitir a un autor autenticarse y administrar su perfil
    - Permitir a un usuario sugerir cambios o pedir la corrección de un dato erróneo
    - Integrar con otros servicios de perfil de autor e identificadores persistentes para el intercambio de metadatos
- Mejorar el proceso de **deduplicación** de los datos de los autores
- Incluir el resto de los autores **UNLP** en el armado de los perfiles.

¡Muchas Gracias!

¿Preguntas?