Detección de registros académicos duplicados obtenidos desde repositorios digitales

Autor Soloaga Ignacio

Directora De Giusti Marisa Raquel

Asesor Profesional Lira Ariel Jorge

Objetivo

- Desarrollar una herramienta de análisis de metadatos para detectar registros duplicados entre repositorios digitales.

Objetivos secundarios

- Definir un flujo de trabajo para la recuperación, limpieza, deduplicación e ingesta masiva de documentos a un repositorio digital.
- Desarrollar una herramienta de mapeo de metadatos para dar soporte a las distintas etapas del proceso.
- Mejorar la cantidad y calidad de metadatos de registros existentes en un repositorio.
- Aumentar la cobertura de obras de autores de la UNLP en el repositorio institucional SEDICI mediante la incorporación de sus obras.
- Permitir encontrar duplicaciones de registros en un repositorio.

Motivación

- Recuperar producción de la UNLP dispersa en la web.
- Aumentar la cantidad y calidad de las publicaciones en SEDICI.
- Asegurar que no se incluirán en importaciones documentos ya existentes en el repositorio.
- Evitar a los autores el autoarchivo en múltiples repositorios.
- Definir un flujo de trabajo que se adapte a cada caso de uso en particular.

Agenda

- Marco teórico
- Deduplicación de registros académicos
- Análisis y desarrollo de la solución
- Desarrollo de aplicación web
- Proceso de importación y casos de aplicación
- Conclusiones y trabajos futuros

Marco teórico

Repositorios digitales

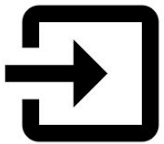
Un **repositorio** es un espacio centralizado donde se **almacena de forma organizada información digital**, en su mayoría producción científica y/o académica de investigadores e instituciones.

SEDICI - Escenario de trabajo

- Repositorio institucional de la UNLP.
- Actualmente alberga más de 100.000 publicaciones.
- Producción científica y académica de docentes e investigadores UNLP.

Ingesta masiva de registros

- Múltiples vías de carga a un repositorio
  - Autoarchivo
  - Vía administración
  - Depósito semi-automático desde servicios de la UNLP.
  - **Importaciones masivas de registros**

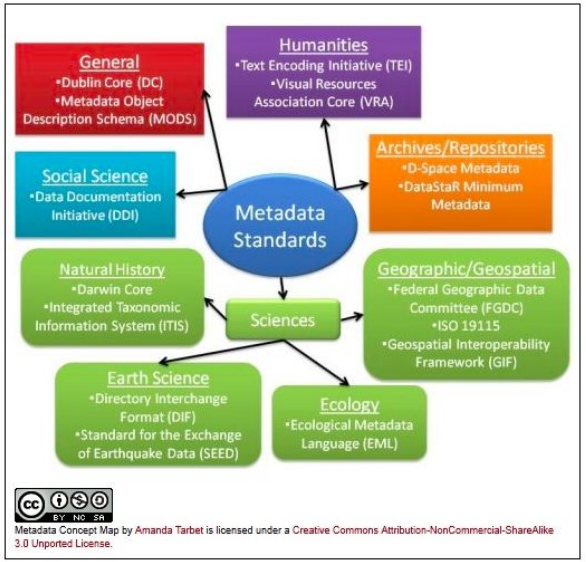


Metadatos

- Describen e identifican a un conjunto de objetos digitales.
- Se organizan bajo **esquemas de metadatos.**
    - **Perfiles de aplicación**
- **Diferentes tipos**: descriptivos, de estructura, técnicos, administrativos.

dc.subject	Turbulence	es
dc.subject	Large eddy simulation	es
dc.subject	Structural vibrations	es
dc.title	Stochastic wind-load model for building vibration estimation using large-eddy cfd simulation and random turbulenc flow generation algorithms	en
dc.type	Objeto de conferencia	es
sedici.identifier.uri	https://cimec.org.ar/ojs/index.php/mc/article/view/5289	es
sedici.identifier.issn	2591-3522	es
sedici.creator.person	Inaudi, José A.	es
sedici.creator.person	Sacco, Carlos G.	es
sedici.description.note	Publicado en: *Mecánica Computacional* vol. XXXV, no. 12	es
sedici.subject.materias	Ingeniería	es

Figura 2. Parte de un registro perteneciente a una ponencia cargada en SEDICI



Registro en lenguaje XML (Dublin Core)

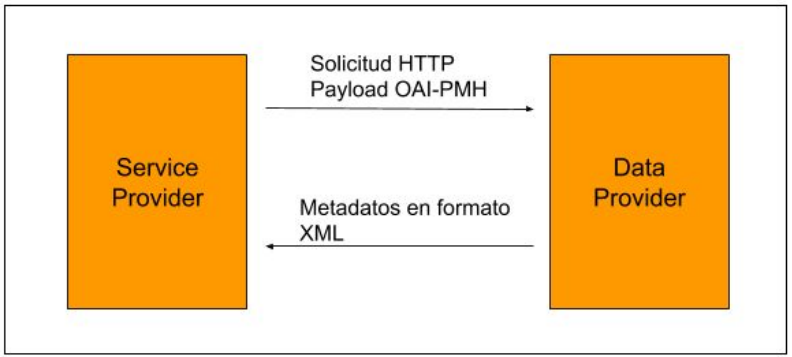
<dc:identifier>**https://doi.org/10.35537/10915/1063**</dc:identifier>
<dc:title>**Ajuste de las variables que gobiernan los modelos de comportamiento de HDM-4 para vías no pavimentadas de la región de Antofagasta (Chile)**</dc:title>
<dc:creator>**Rojas Cazaluade, Oscar Orlando**</dc:creator>
<dc:date>**2008**</dc:date>
<dc:contributor>**Campana, Juan Manuel**</dc:contributor>
<dc:contributor>**Infante, José Luis**</dc:contributor>
<dc:language>**es**</dc:language>
<dc:subject>**Ingeniería**</dc:subject>
<dc:subject>**ingeniería vial; carreteras; modelos de deterioro y conservación de caminos; pavimentos; rugosidad**</dc:subject>
<dc:subject>**Antofagasta (Chile)**</dc:subject>

Interoperabilidad

- **Entre sistemas**: colaboración a través de la comunicación.
- **Entre esquemas de metadatos**: la información puede ser expresada de igual manera en dos esquemas de metadatos distintos sin pérdida de información.

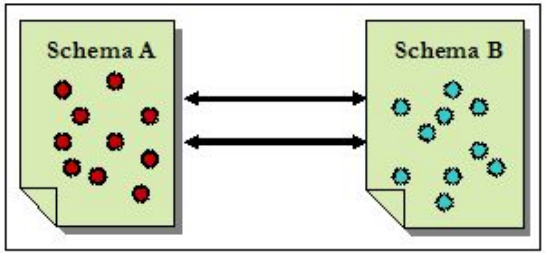
Protocolo **OAI-PMH**

- Protocolo para la **cosecha de metadatos** de cualquier material en soporte digital
- **Dublin Core**
- **Service** Provider y **Data** Provider



Mapeo de metadatos

Proceso que consiste en analizar dos esquemas de metadatos para **encontrar equivalencias** entre los mismos.



Necesidad de mapear:

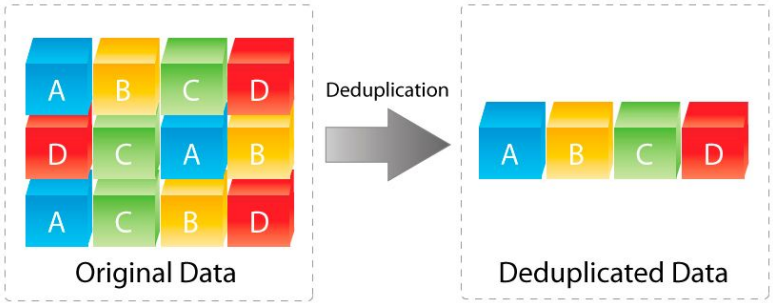
- **Elementos** (dc.author -> sedici.creator.person)
- **Semántica** (significado)
- **Sintaxis** (estructura)

Detección de registros académicos duplicados

Deduplicación de registros

Encontrar **registros** en un conjunto de datos que **hacen referencia a la misma entidad** en distintas fuentes de datos.

- **Distintos mecanismos** según el dominio del problema.
- Importancia de la **estructura de los datos** (heterogeneidad)
- Tres grandes etapas:
  - Parseo y estandarización
  - Transformación (mapeo)
  - Coincidencia de registros



Deduplicación de registros académicos

- Formatos y **esquemas variados** con **estructuras diferentes**
- Se deben analizar los metadatos de cada registro.
    - Utilizando **funciones de distancia** entre *strings*.
- Deduplicación **interna versus cruzada**.
- **Heterogeneidad en los metadatos** de distintas fuentes.
- Uso de **identificadores** persistentes **muy bajo**.

Heterogeneidad de los datos

- Titulo: ‘Políticas territoriales y construcción del paisaje cultural’
- Subtitulo: ‘Caso Región Gran La Plata’
- Políticas territoriales y construcción del paisaje cultural - Caso región gran La Plata

- ‘García, María Ana’
  - ‘García, María A.’
  - ‘García, M. Ana’
  - ‘García, M. A.’

- Fecha de publicación ≠ fecha de exposición ≠ fecha de depósito.
- Bajo porcentaje de identificadores persistentes asociados a cada registro.

Técnicas para la detección de registros duplicados

- Modelos probabilísticos de emparejamiento
- Aprendizaje supervisado y semi supervisado
- Técnicas basadas en aprendizaje activo
- Técnicas basadas en distancia
- Enfoques basados en reglas
- Aprendizaje sin supervisión

También existen **metodologías para optimizar cantidad de comparaciones**:

- Métodos de bloques
- Métodos de tipo ventana

Soluciones existentes

OpenAIRE

Dedupe.io

Buscadores y gestores de ref. bibliográfica

(Mendeley, Ovid, Refworks, Endnote, Google Scholar)

MENDELEY

Soluciones extra-dominio

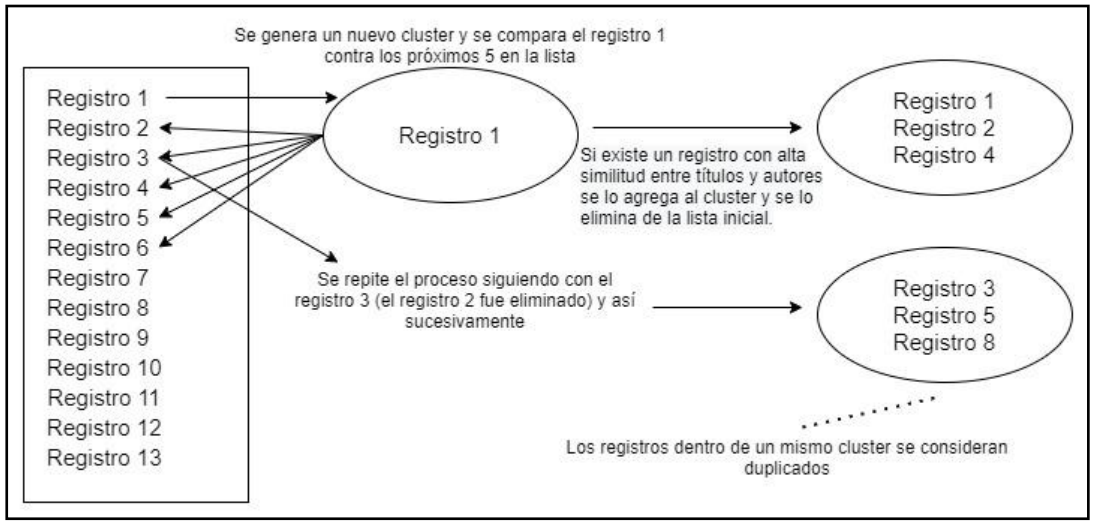
y soluciones comerciales

Análisis y desarrollo de la solución

Un primer prototipo ...

- Compara sólo título y autores
- Utiliza algoritmo de tipo ventana (**clave** = **título**)
- Problemas e insuficiencias

Surge la necesidad de modelar una herramienta más robusta.



Solución propuesta

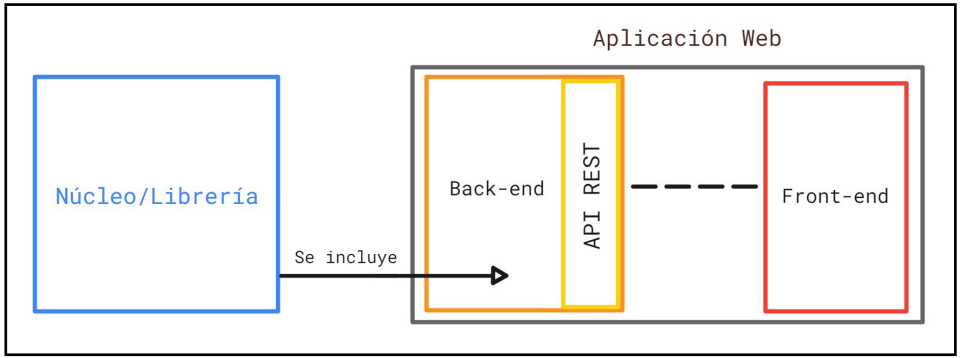
Se propone el desarrollo de un **sistema basado en reglas y funciones de distancia** que permita:

- **Comparar** dos conjuntos de registros de metadatos **en busca de duplicados**.
- Establecer **porcentajes de similitud** entre cada par de registros comparados.
- **Umbrales configurables** a partir de los cuales dos registros se consideran duplicados, casi duplicados, no duplicados o indefinido.
- **Flexibilidad** para incorporar **nuevos formatos de archivo**.
- **Flexibilidad** para comparar **nuevas tipologías** de registros.

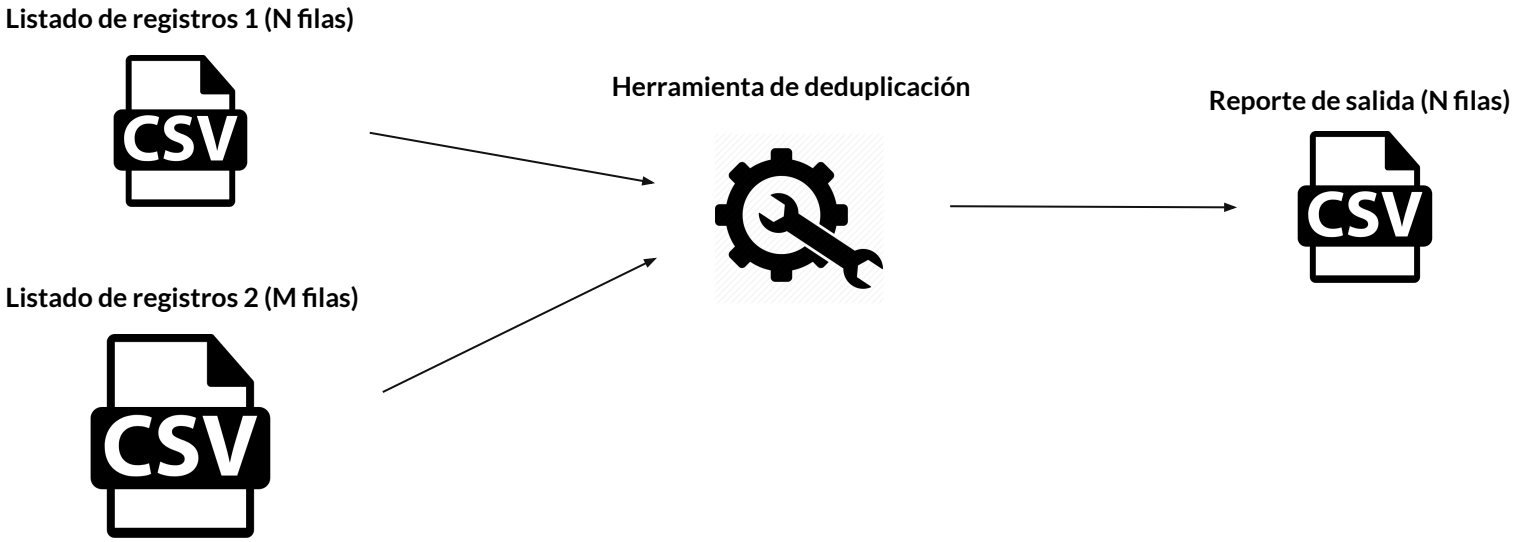
Arquitectura de la herramienta

División en dos componentes: **librería** y **aplicación web**.

- **Librería** (Nucleo): concentra lógica principal de la herramienta.
- **Aplicación web:** permite el uso de la herramienta a través de una interfaz de usuario.



Funcionamiento general



Esquema de metadatos genérico

Necesidad de almacenar y acceder a la información de cada registro de manera uniforme.

Normalización de la tipología de cada registro

Ej.: 'info:eu-repo/semantics/article' vs 'Article' vs 'Articulo'

Se define una **función encargada de normalizar la tipología** de un registro que se basa en un listado de posibles valores para cada tipo de documento.

Metadato	Condición
id	Obligatorio
title	Obligatorio
subtitle	Opcional
type	Obligatorio
author	Obligatorio
date	Obligatorio
doi	Opcional
isbn	Opcional
issn	Opcional
description	Opcional

Engine y algoritmo de comparación

Clase Engine: encargada de gestionar el proceso de deduplicación de principio a fin.

La gestión de resultados asociados a cada regla y escritura del reporte de salida es delegada a las clases **ResultsCollector**, **OutputFormatter** y **OutputHandler** respectivamente.

```
for registroN in listado_de_registrosN:
    for registroM in listado_de_registrosM:
        Seleccionar conjunto de reglas en base al registroN y registroM.
        for regla1 in conjunto_de_reglas:
            Evaluar regla sobre el registroN y registroM
            if resultado > 0.5:
                Agregar resultado al listado de resultados del registroM
        Determinar si registroM es candidato a duplicado
    Seleccionar candidatos a duplicados para el registroN
    Generar tupla de salida para el registroN
```

Resultado ejemplo para un registro

```
id_documentN: '10915/91176',
ids_documentsM: '11746/10511',
rules: {
            regla1: 1,
            regla2: 0.75
},
similarity: 'DUPLICATE'
```

Sistema de reglas

Cada **regla** evalúa la similitud existente entre un **subconjunto limitado de los metadatos** de cada registro.

La evaluación de una regla sobre un par de registros genera como resultado un objeto **Result**, el cual contiene información sobre el nombre de la regla que lo produjo, el puntaje calculado y si la regla pudo o no votar.

Con **niveles de aceptación predefinidos** por cada regla, se determina un **porcentaje de similitud** general para cada par de registros en particular.

Valor	Significado
1	Registros duplicados
0.75	Registros casi duplicados
0.5	Indefinido
0.25	(No se utiliza)
0	Registros no duplicados

Posibles valores generados por una regla

Tipos de reglas

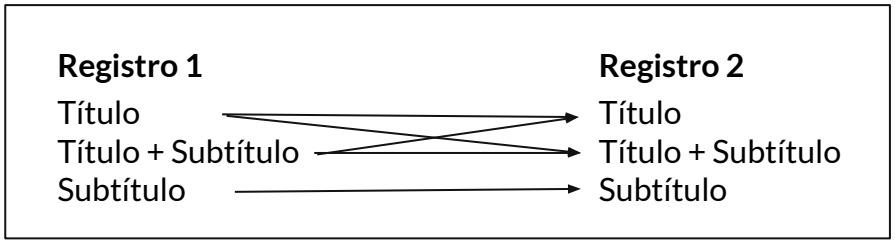
Se definieron reglas específicas que evalúan pares de documentos **en base a la tipología** de los mismos, y que permiten que cada una pueda centrarse en el **análisis de metadatos específicos según el tipo de documento** a comparar.

Nombre de la clase	Aplicabilidad	Metadatos utilizados
GeneralRule	Todos	Título, autores y fecha
JournalArticleRule	Artículos	ISSN, título, autores y fecha
BookRule	Libros	ISBN y título
ThesisRule	Tesis	Título y autor/es
BookChapterRule	Capítulos de libro	ISBN, título del libro, título del capítulo
DoiRule	Todos	DOI y título
AbstractRule	Todos	Resumen o abstract

Comparación de títulos

Dificultades

- Título y subtítulo guardados en el mismo metadato
- Títulos en distintos idiomas
- Errores de tipeo
- Diferencias en la estructura
- Presencia de información adicional



Además se contemplan situaciones en las que un registro posee el título en varios idiomas.

Comparación de autores

Nombre autor 1	Nombre autor 2	Nivel de coincidencia
García, Juan	García, J.	Alto
Fernández García, Juan	Fernández G., Juan	Alto
Perez, J.	Perez García, J.	Alto
de la Paz Diulio, María	Diulio, María de la Paz	Bajo
Perez, J.	Peres, J.	No hay coincidencia
García, Gabriel	García, María	No hay coincidencia
Fernández, Alfredo Horacio	Fernández, Horacio	Bajo

Comparación de fechas

- **‘05-11-2018’** -> día **5 del mes de noviembre** o al **día 11 del mes de mayo**.

  Necesitamos información sobre el formato utilizado **(DD-MM-AAAA, MM-DD-AAAA).**

- **Múltiples fechas asociadas a cada registro** (preprints, postprints y versiones publicadas)

Entonces…

- Se genera un puntaje para determinar si dos registros tienen **fechas de publicación iguales** (1) o no las tienen (0).
- Compara únicamente los **años de cada fecha**.
- Realiza una comparación de **todas contra todas** (en caso que se cuente con más de una fecha en alguno de los dos registros a comparar).

Módulo auxiliar *utils*

- Calcular la **similitud** entre dos strings.
- Obtener el **ISSN** de un registro.
- Obtener el **ISBN** de un registro.
- Obtener el **DOI** de un registro.
- **Normalizar** un string
- **Normalizar el tipo** de un registro.

Funciones de similitud entre strings analizadas

Distancia Levenshtein

Calcula el número de **inserciones, eliminaciones o sustituciones** de caracteres necesarias para transformar un **string A** en un **string B**.

hola | ola
hola | **h**ola

Distancia = 1

Distancia Jaro-Winkler

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

$$sim_w = sim_j + \ell p(1 - sim_j),$$

Metaphone

- Algoritmo fonético.
- Indexa palabras en base a su pronunciación en inglés.
- Genera claves. Ejemplo: (‘LSPRSLTRRS’) ó (‘APRSLTRRS’)

Función de similitud entre strings utilizada

Se definió una función que retorna un porcentaje de similitud entre dos strings en base a la relación entre la **a) distancia *Levenshtein* y b) la longitud máxima entre los mismos.**

```
def calculateDistance(string1, string2):
      try:
          text_distance = Levenshtein.distance(string1, string2)
          max_len = max(len(string1),len(string2))
          normalized_distance = 1 - text_distance / max_len
      except:
          normalized_distance = 0
      return normalized_distance
```

Resultado de una deduplicación

Registros a deduplicar	Registros duplicados	Reglas	Similitud
https://ri.conicet.gov.ar/handle/11336/14093	http://sedici.unlp.edu.ar/handle/10915/36493	JournalArticleRule=1 | GeneralRule=0.75 | DoiRule=A	DUPLICATE
https://ri.conicet.gov.ar/handle/11336/69795	http://sedici.unlp.edu.ar/handle/10915/53734	JournalArticleRule=1 | GeneralRule=0.75 | DoiRule=A	DUPLICATE
https://ri.conicet.gov.ar/handle/11336/9668	http://sedici.unlp.edu.ar/handle/10915/72967	GeneralRule=0.75 | DoiRule=A|	NEAR_DUPLICATE
https://ri.conicet.gov.ar/handle/11336/69595	None	None	NO_DUPLICATE

Artículo

Comunicación y salud en América Latina : un panorama de las perspectivas, los itinerarios teórico-prácticos y los desafíos actuales

Bruno, Daniela; Demonte, Flavia Carolina

Fecha de publicación: 09/2015
Editorial: Universidad Nacional de La Plata. Facultad de Periodismo y Comunicación Social
Revista: Actas de Periodismo y Comunicación
ISSN: 2469-0910
Idioma: Español
Tipo de recurso: Artículo publicado

Resumen

El campo de la comunicación en salud se constituye como lugar de reflexión académica en EEUU y algunos países europeos durante las décadas de 1960 y 1970 en el contexto de los programas de desarrollo internacional, en especial, los preocupados por el control demográfico, la planificación familiar y la educación sanitaria (Petracci y Waisbord, 2011). Por ende, sus orígenes están marcados tanto por las bases conceptuales y operativas de las teorías modernistas y los lineamientos de las políticas pensadas en ese contexto, basados en el principio conductista del cambio de comportamiento a partir de la exposición reiterada a ciertos mensajes. No obstanteesta hegemonía difusionista de los inicios, es importante tener en cuenta que muy tempranamente comenzaron a registrarse en el continente latinoamericanoexperiencias de resistencia y crítica al difusionismo, bajo el paraguas de lo que hoy conocemos como comunicación popular, alternativa y comunitaria.

Palabras clave: Comunicación , Salud , Enfoques , América Latina

Ver el registro completo

Comunicación y salud en América Latina

Un panorama de las perspectivas, los itinerarios teórico-prácticos y los desafíos actuales

Autores: Bruno, Daniela Paola | Demonte, Flavia Carolina **2015**

Tipo de documento: Artículo

Like Share Twittear

Resumen

El campo de la comunicación en salud se constituye como lugar de reflexión académica en EE. UU. y algunos países europeos durante las décadas de 1960 y 1970 en el contexto de los programas de desarrollo internacional, en especial, los preocupados por el control demográfico, la planificación familiar y la educación sanitaria (Petracci y Waisbord, 2011). Por ende, sus orígenes están marcados tanto por las bases conceptuales y operativas de las teorías modernistas y los lineamientos de las políticas pensadas en ese contexto, basados en el principio conductista del cambio de comportamiento a partir de la exposición reiterada a ciertos mensajes. No obstante esta hegemonía difusionista de los inicios, es importante tener en cuenta que muy tempranamente comenzaron a registrarse en el continente latinoamericano experiencias de resistencia y crítica al difusionismo, bajo el paraguas de lo que hoy conocemos como comunicación popular, alternativa y comunitaria.

Notas

Eje temático: Dimensiones sociales y comunicacionales de la salud

Información general

Fecha de publicación: 2015

Idioma del documento: Español

Revista: Actas de Periodismo y Comunicación; vol. 1, no. 1

Evento: II COMCIS y I CCP (La Plata, 2015)

Institución de origen: Facultad de Periodismo y Comunicación Social

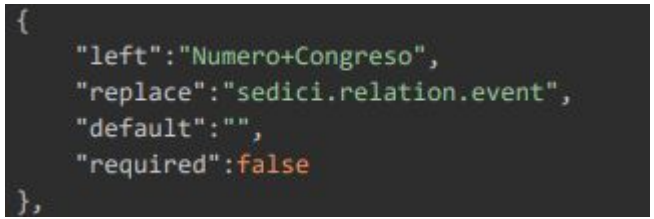
ISSN: 2469-0910

Palabras claves: América Latina ; Salud

Materias: Comunicación

Herramienta para el mapeo de metadatos

- Necesidad constante de realizar mapeos entre esquemas de metadatos/perfiles de aplicación.
- **Configuraciones reutilizables** y customizables en formato JSON.
- Soporte para expresiones regulares.
- **Facilidad para eliminar** gran cantidad de **columnas** automáticamente.
- Posibilidad de **aplicar filtros.**
- Soporte formatos CSV, TSV, etc.



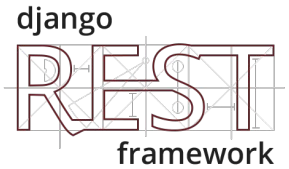
Desarrollo de Aplicación Web

Aplicación back-end

Tecnologías utilizadas

PYTHON

django

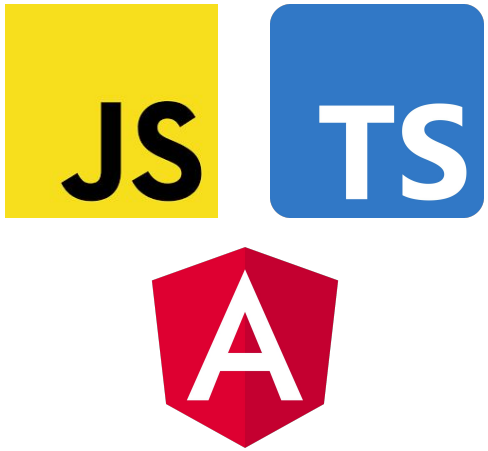


- Funcionalidad expuesta mediante **API REST**
- Extensiones al modelo: **tareas de mapeo** y **tareas de deduplicación**.

Descripción	SEDICI vs Ponencias Memoria Académica
Estado	En progreso
Progreso	75,5 %
Archivo CSV 1	sedici-formato-generico.csv
Archivo CSV 2	ponencias-mem-academica-generico.csv
Resultado	(Vacío)

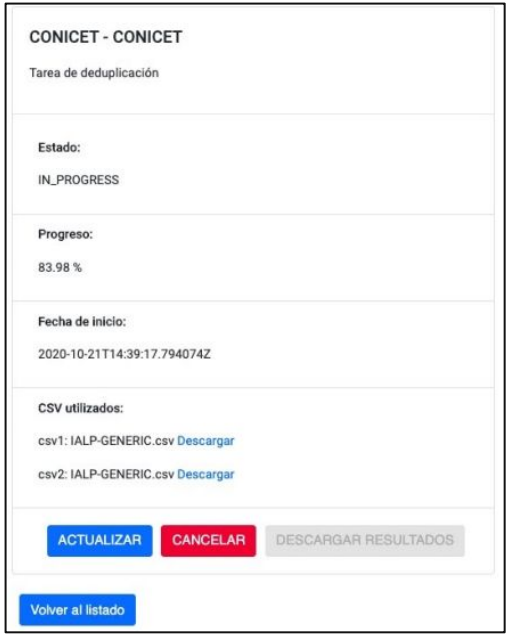
Aplicación front-end

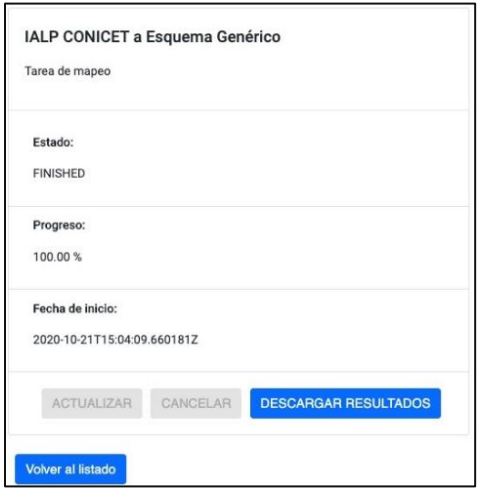
Tecnologías utilizadas



- Interfaz de usuario más accesible.
- Documentación auto-contenida en la herramienta.
- CRUD de tareas de mapeo y deduplicación.
- Extensible a nuevos módulos.

Pantallas principales





Documentación de la herramienta

Deduplicador

La herramienta de deduplicación permite iniciar tareas para deduplicar dos listados de registros. Cada tarea iniciada se guarda en el sistema con una descripción y un estado que se actualiza con el progreso de la tarea. El listado de registros debe subirse al servidor en archivos con formato CSV, donde las columnas de cada archivo representan cada metadato del registro. Por ejemplo el título, el autor o las fechas de la obra. Estos archivos se guardan en el sistema asociados a cada tarea para poder ser descargados si se desea.

Guía de uso

Para inciar una tarea de deduplicación debe mover el cursor hacia la opción Deduplicator en el menú de navegación y hacer click en la opcion Iniciar tarea. Luego, debe completar el formulario indicando:

- Descripción de la deduplicación
- CSV con registros 1
- CSV con registros 2
- Multithread

Formato archivos CSV

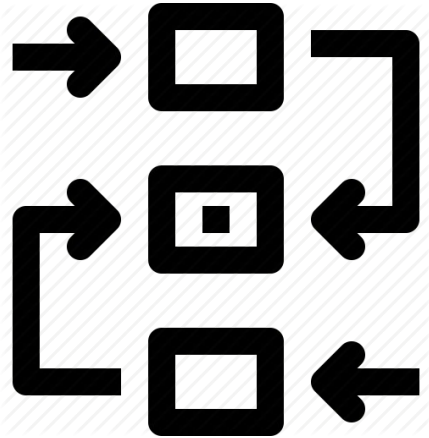
Cada celda del CSV debe contener el valor correspondiente al metadato indicado en el nombre de la columna. Para campos multivaluados se debe utilizar el caracter '**|**' como separador de valores. Ej. 'Fernández, Pablo|Almirante, Roberto|Giménez, Graciela'.

Proceso de importación

Casos de aplicación

Definición del proceso

1. Obtención de registros desde un repositorio
2. Mapeo de metadatos a formato genérico
3. Deduplicación con registros del repositorio destino
4. Reconciliación de metadatos
5. Mapeo a formato esperado por el repositorio destino
6. Correcciones sobre los metadatos
7. Obtención de los objetos digitales asociados a cada registro
8. Generar archivo de importación y carga del mismo



Casos de aplicación

Memoria Académica

Se obtuvieron y procesaron:

- **9.644** registros de artículos
- **529** registros de libros
- **13.615** registros de eventos

Se importaron:

- 1.491 registros de artículos
- 310 registros de libros
- 2.643 registros de eventos

CONICET Digital

Se obtuvieron y procesaron:

- **16.153** registros de artículos.

Fuente de obtención	Cantidad de registros
Colección OAI 'CCT La Plata'	12476
Registros con filiación 'UNLP'	1595
Registros que pertenecen a colecciones de centros UNLP	1393

Se importaron:

- 3.366 registros de artículos

SCOPUS

Se obtuvieron y procesaron:

- **4.260** registros de artículos UNLP y AA.

Se importaron:

- 3.309 registros de artículos

Conclusiones y trabajos futuros

Conclusión

- Investigación de técnicas y metodologías para la deduplicación de registros.
- Implementación de una herramienta de detección de registros duplicados.
  - Se procesaron alrededor de 150.000 registros.
- Incorporación de 11.000 documentos nuevos al repositorio SEDICI.
- Disminución considerable del tiempo de carga de las publicaciones.
- Definición de un proceso que puede servir a múltiples repositorios y distintos casos de uso.

Trabajos futuros

- Mejorar performance de la herramienta de deduplicación
- Expandir módulo de comparación de autores
- Enriquecimiento de registros detectados como duplicados
- Explorar enfoque de Aprendizaje Automático
- Incorporar funcionalidad de deduplicación dentro del sistema de repositorio

¡Gracias!

¿Preguntas?

Soloaga Ignacio