“Herramientas para la Interoperabilidad y Normalización de datos en RI”

Tesina Licenciatura en Sistemas

FACULTAD DE INFORMÁTICA

UNLP

Autor: Almazán, María Belén
Director: Ing. De Giusti, Marisa

Motivación

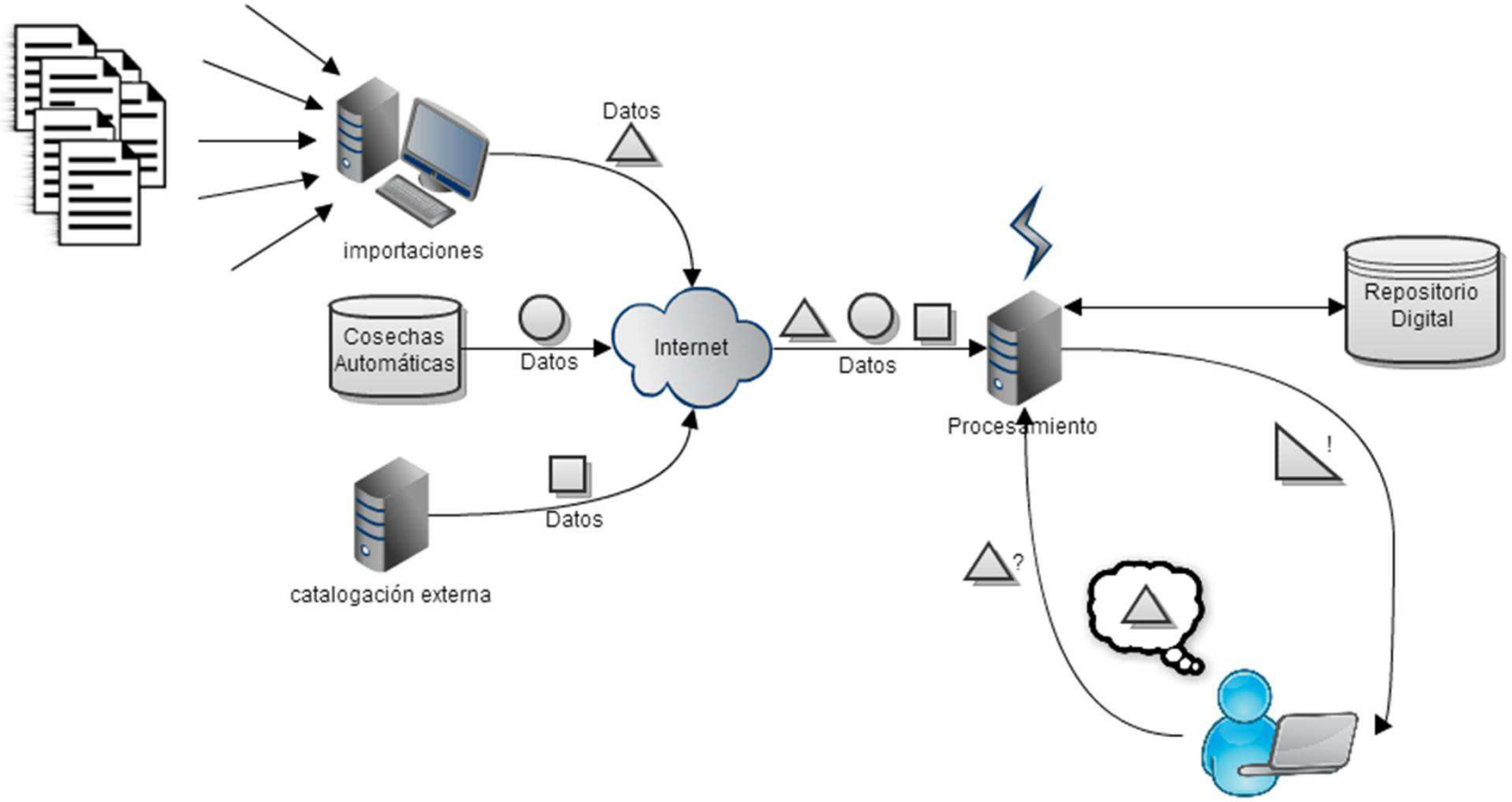

Figura 1: Heterogeneidad de la Información

Objetivos

- Desarrollar métodos de depuración, asociación, inferencia y normalización, que mejoren la calidad de los datos pertenecientes a un repositorio institucional de modo que se pueda explotar al máximo la información allí contenida.

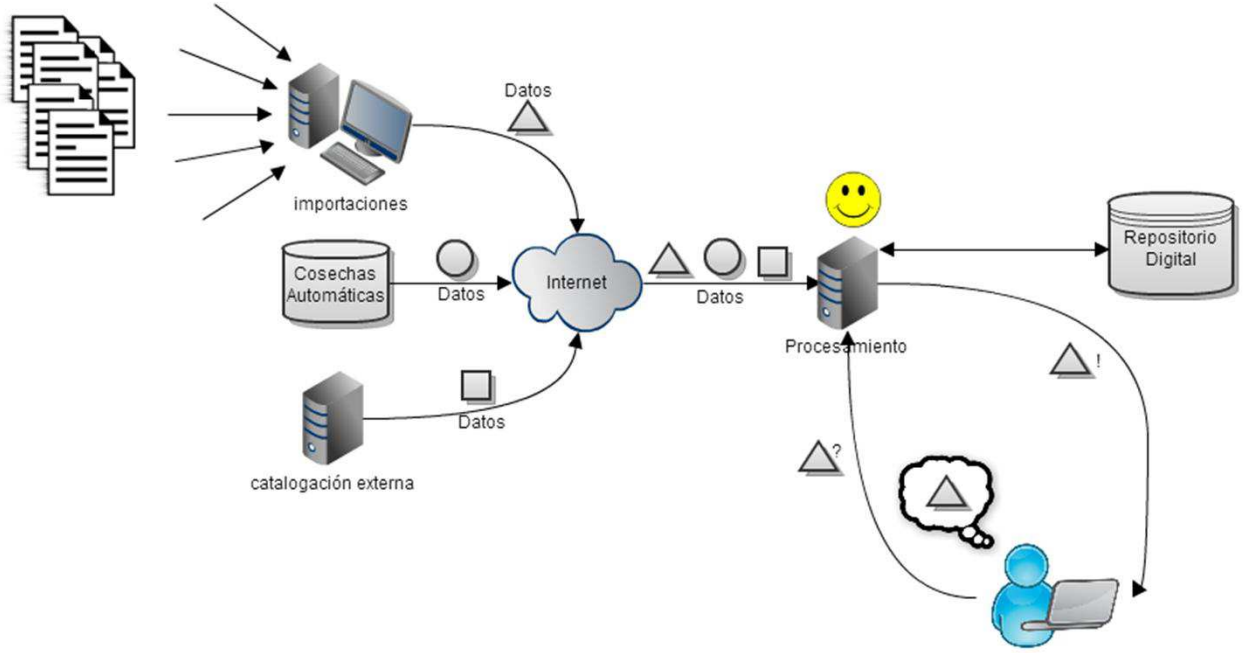

Figura 2: Información Normalizada

- **¿Qué es un repositorio Institucional?**
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
- Herramienta de Cosecha

¿Qué es un Repositorio Institucional?

- Un *repositorio* es una infraestructura web capaz de brindar un conjunto de servicios a una comunidad, destinados a recopilar, gestionar, difundir y preservar contenidos a través de una colección organizada y accesible en abierto que debe estar provista de facilidades que le permiten interoperar con otros repositorios similares.

- Un *repositorio institucional*, en particular, contiene, como su nombre lo indica, la producción intelectual de cierta comunidad académica

- ¿Qué es un repositorio Institucional?
- **Metadatos**
- Acceso Abierto
- Interoperabilidad
- Normalización
- Herramienta de Cosecha

Metadatos

- Son datos estructurados que describen otros datos; en otras palabras, se trata de datos sobre datos.

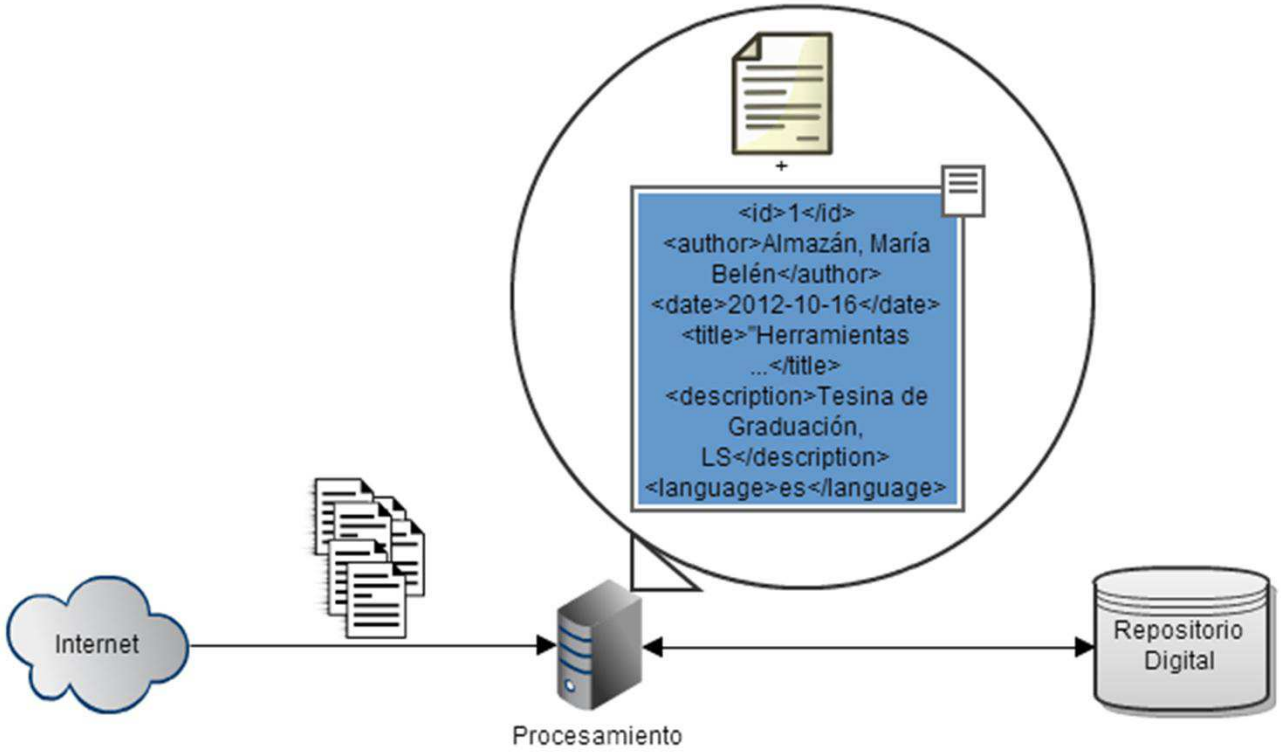

Figura 3: Metadatos

- ¿Qué es un repositorio Institucional?
- Metadatos
- **Acceso Abierto**
- Interoperabilidad
- Normalización
- Herramienta de Cosecha

Acceso Abierto

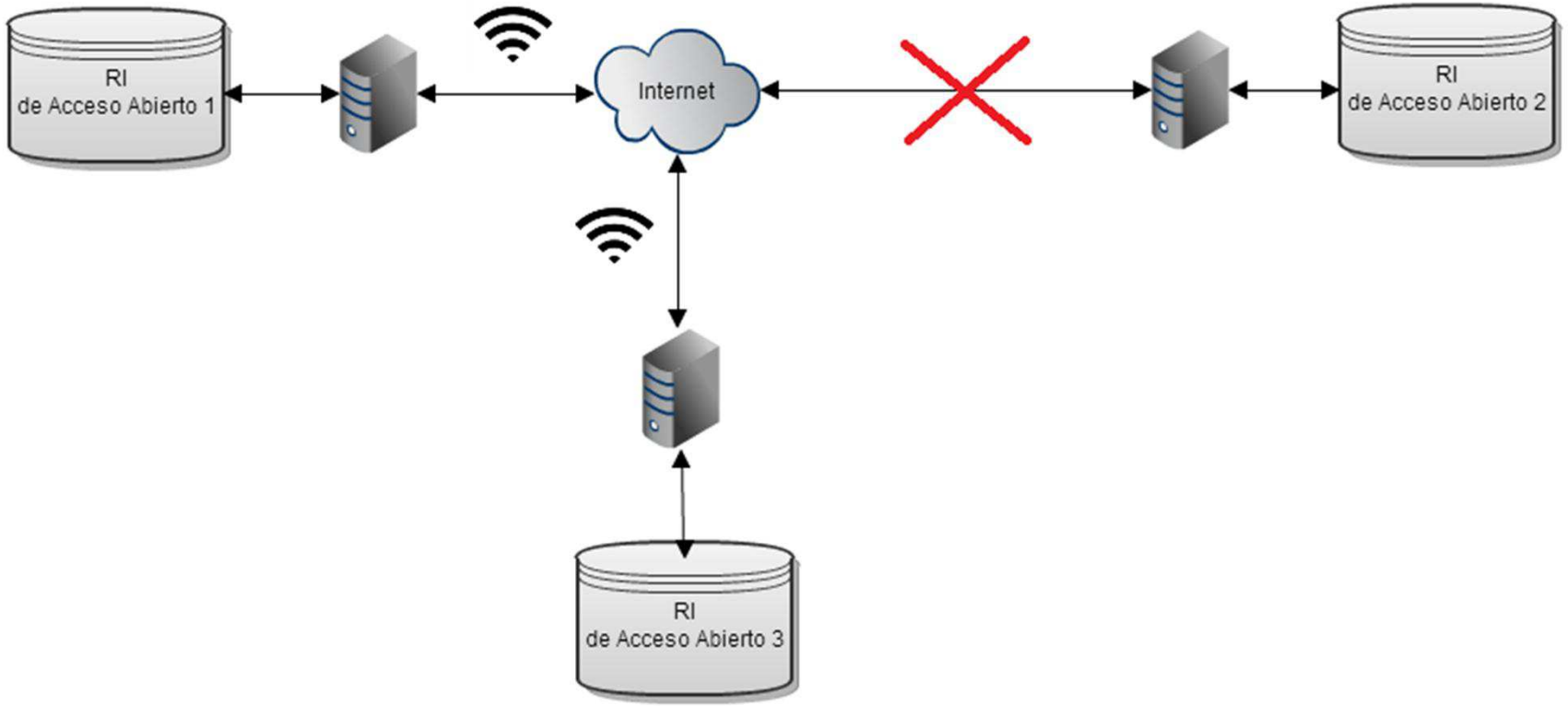

Figura 4: Acceso Abierto – Diferencias de plataformas o tecnologías

→Es necesario que exista *interoperabilidad* entre los repositorios

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- **Interoperabilidad**
- Normalización
- Herramienta de Cosecha

Interoperabilidad

- Capacidad para comunicar sistemas entre ellos e intercambiar información en un formato utilizable.

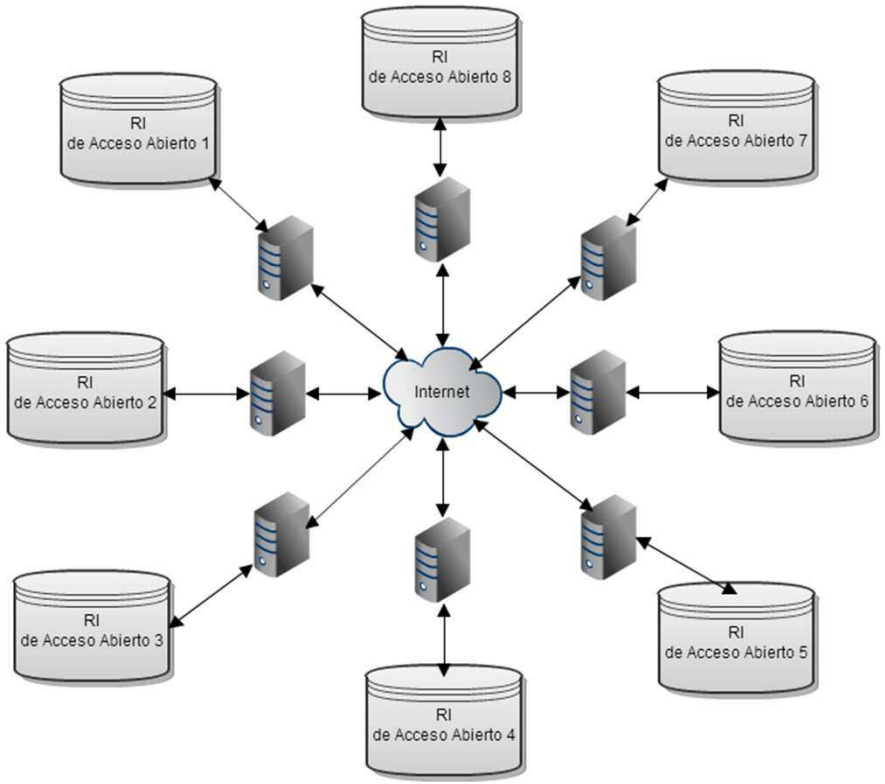

Figura 5: Interoperabilidad

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- **Normalización**
  - Control de Calidad
  - Enriquecimiento de la Información
- Herramienta de Cosecha

Normalización

- Para optimizar las técnicas de recuperación de la información e interoperabilidad, es necesario mejorar la calidad de los metadatos de las bases de datos documentales.

→Se utilizan procesos de normalización

- Objetivo:
  - Minimizar redundancias
  - Simplificar el mantenimiento de los datos
  - Permitir la recuperación sencilla de los datos
  - Evitar datos no identificables

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
  - **Control de Calidad**
    - Inconsistencias
    - Datos Incompletos
    - Anomalías
  - Enriquecimiento de la Información
- Herramienta de Cosecha

Control de Calidad

- Consiste en detectar errores en la información y luego corregirlos

- Se tienen en cuenta 3 tipos de errores:
  - Inconsistencias
  - Datos Incompletos
  - Anomalías

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
  - Control de Calidad
    - **Inconsistencias**
    - Datos Incompletos
    - Anomalías
  - Enriquecimiento de la Información
- Herramienta de Cosecha

...Inconsistencias

- Detección de registros que no cumplen con determinadas reglas, y posterior modificación de los datos para que sí las cumplan.

- Esta tarea incluye asegurar que la información se encuentre consistente y libre de redundancias

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
  - Control de Calidad
    - Inconsistencias
    - **Datos Incompletos**
    - Anomalías
  - Enriquecimiento de la Información
- Herramienta de Cosecha

...Datos Incompletos

- Ejemplo: valores nulos en una base de datos relacional

- Dos tipos de fuentes de incompletitud:
  - Datos truncados
  - Datos censurados

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
  - Control de Calidad
    - Inconsistencias
    - Datos Incompletos
    - **Anomalías**
  - Enriquecimiento de la Información
- Herramienta de Cosecha

...Anomalías

- Cuando el valor de un dato difiere en gran medida con respecto a los demás datos.

- Situaciones:
  - El valor fue mal medido, o mal ingresado en la base.
  - El valor corresponde a una "muestra" distinta a la de todos los demás.
  - El valor es correcto y simplemente corresponde a algún suceso inusual de la realidad.

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
  - Control de Calidad
    - Inconsistencias
    - Datos Incompletos
    - Anomalías
  - **Enriquecimiento de la Información**
- Herramienta de Cosecha

Enriquecimiento de la Información

- Completar los datos almacenados en un repositorio con información más precisa

- Añadir información inexistente, de buena calidad, proveniente de distintas fuentes de datos, para incrementar la potencialidad de los datos residentes en el repositorio local.

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
- **Herramienta de Cosecha**
  - Harvester-Arquitectura
  - Harvester-Filtros Existentes
  - Harvester-Problema Existente

Herramienta de Cosecha

Con el objetivo de:

- Recolectar datos desde diferentes repositorios
- Maximizar la cantidad de documentos ofrecidos
- Minimizar el esfuerzo de procesamiento y conexión que implica la tarea de recolección

SeDiCI desarrolló una herramienta general de recolección de recursos denominada “Harvester”

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
- Herramienta de Cosecha
  - **Harvester-Arquitectura**
  - Harvester-Filtros Existentes
  - Harvester-Problema Existente

Harvester - Arquitectura

- Sigue el patrón de Arquitectura ETL

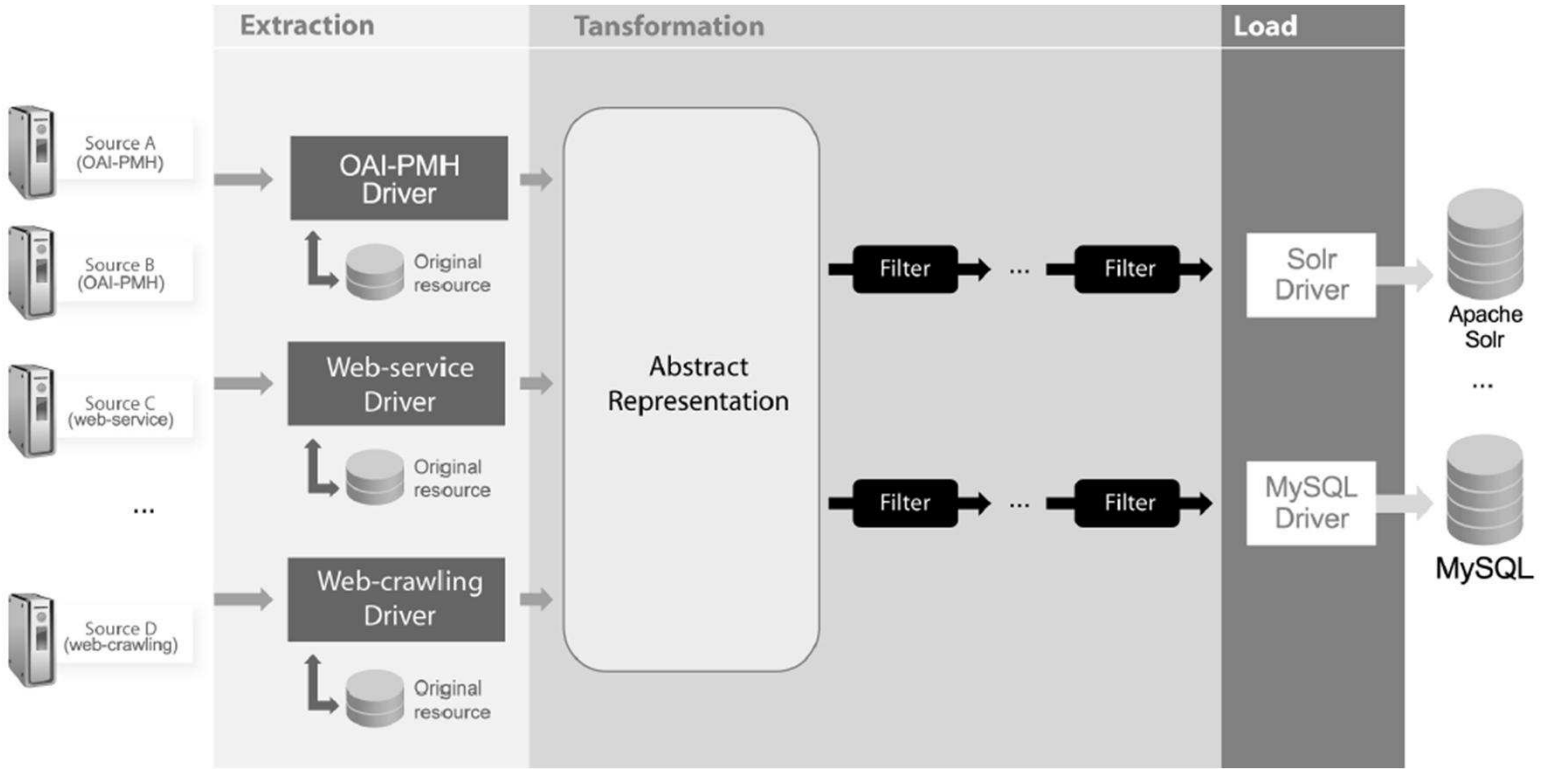

Figura 6: Harvester - Diagrama de Arquitectura

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
- Herramienta de Cosecha
  - Harvester-Arquitectura
  - **Harvester-Filtros Existentes**
  - Harvester-Problema Existente

Harvester – Filtros Existentes

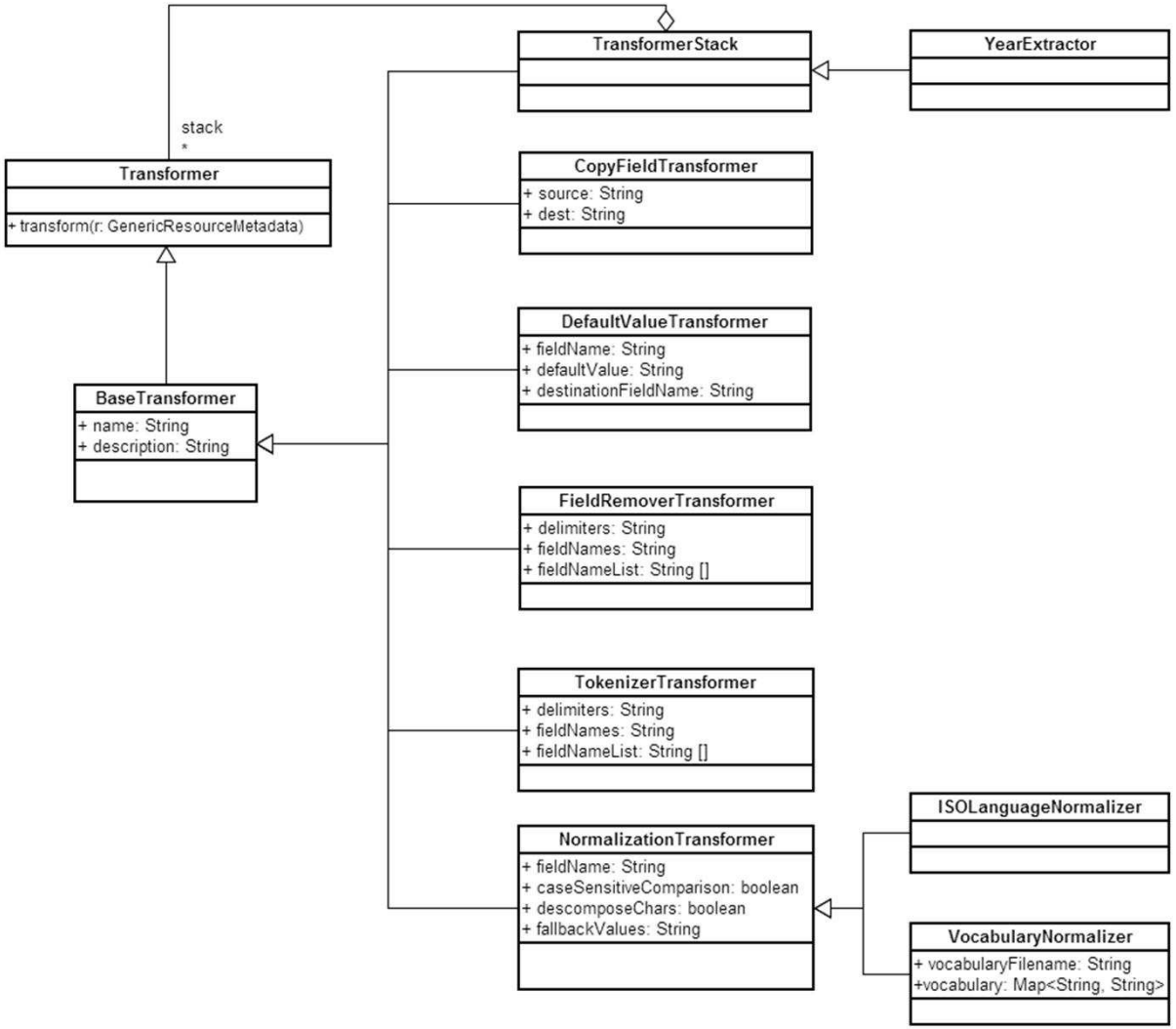

Figura 7: Harvester – Filtros Existentes

- ¿Qué es un repositorio Institucional?
- Metadatos
- Acceso Abierto
- Interoperabilidad
- Normalización
- Herramienta de Cosecha
  - Harvester-Arquitectura
  - Harvester-Filtros Existentes
  - **Harvester-Problema Existente**

Harvester – Problema Existente

- Recursos bibliográficos pertenecientes a una gran diversidad de fuentes → gran heterogeneidad de datos.
- Dificultad para procesar e intercambiar datos.
- Aplicación de transformaciones → posible pérdida de información o generación de documentos incompletos

Es necesario normalizar los datos

- **Aporte**
- Metodología
- Filtros Desarrollados
- Conclusiones
- Trabajos Futuros

Aporte

Se estudiaron las necesidades de normalización existentes en SeDiCI:

- Valores del texto de un mismo metadato con formatos distintos
- Incompletitud del metadato idioma
- Nombre de un mismo autor citado de modos diversos
- Heterogeneidad de formatos de fecha

...Aporte

- ✓ Se implementaron una serie de filtros que generan datos normalizados
- ✓ Se ejecutan en la etapa de Transformación de datos del Harvester
- ✓ Se apuntó a dos áreas:
  - Control de calidad de la información: detección de errores y posteriores correcciones sintácticas.
  - Enriquecimiento de la información

- Aporte
- **Metodología**
- Filtros Desarrollados
- Conclusiones
- Trabajos Futuros

Metodología

- Cambios graduales y/o evolutivos
- Una clase Java para cada Filtro
- Parámetros definidos en archivo de configuración. Ejemplo:

```
<entry key="author_uc">
    <bean
    class="ar.edu.unlp.sedici.harvester.transformers.trans.UpperCaseFieldTransfor
    mer">
        <property name="fieldName" value="author"/>
        <property name="name" value="author_uc"/>
        <property name="makeBackup" value="true"/>
    </bean>
</entry>
```

...Metodología

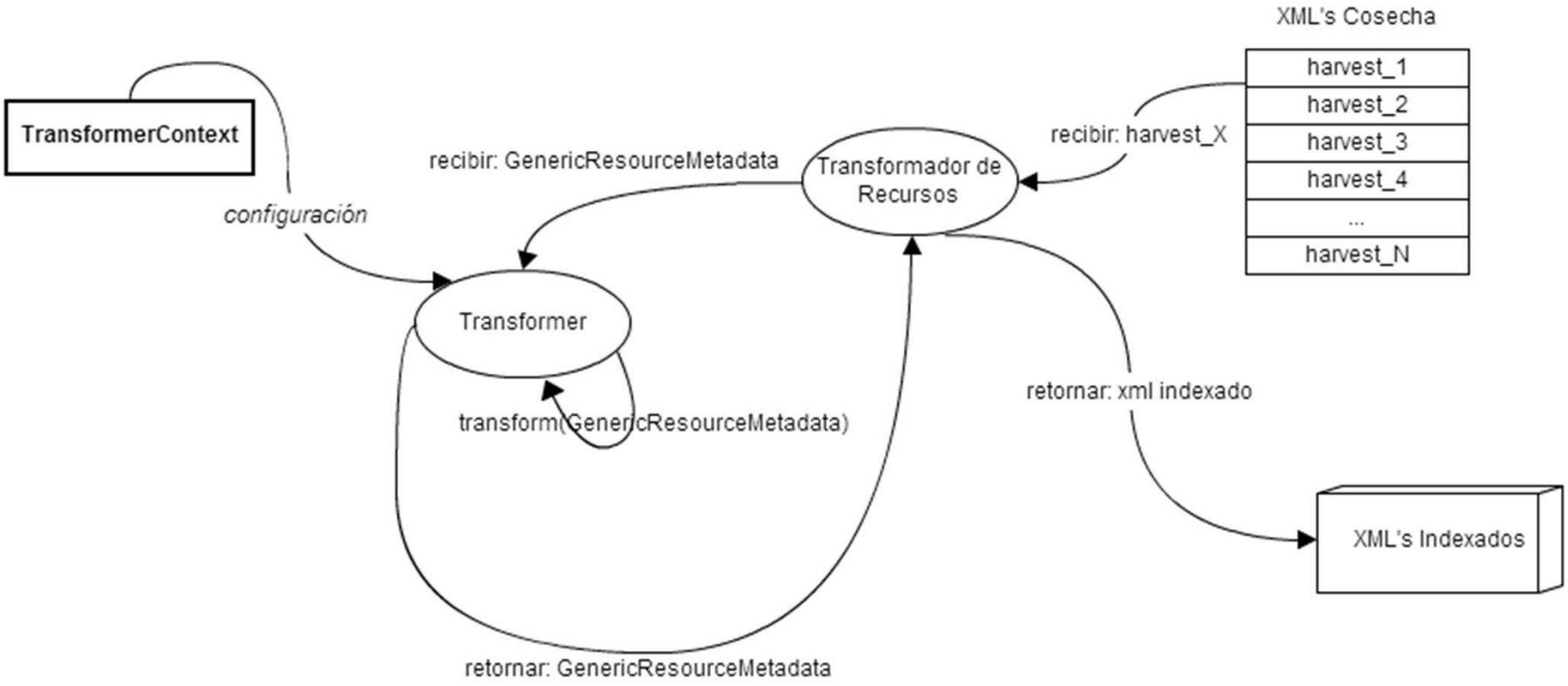

Figura 8: Harvester – Indexación de una Cosecha

- Aporte
- Metodología
- **Filtros Desarrollados**
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

Filtros de Desarrollados

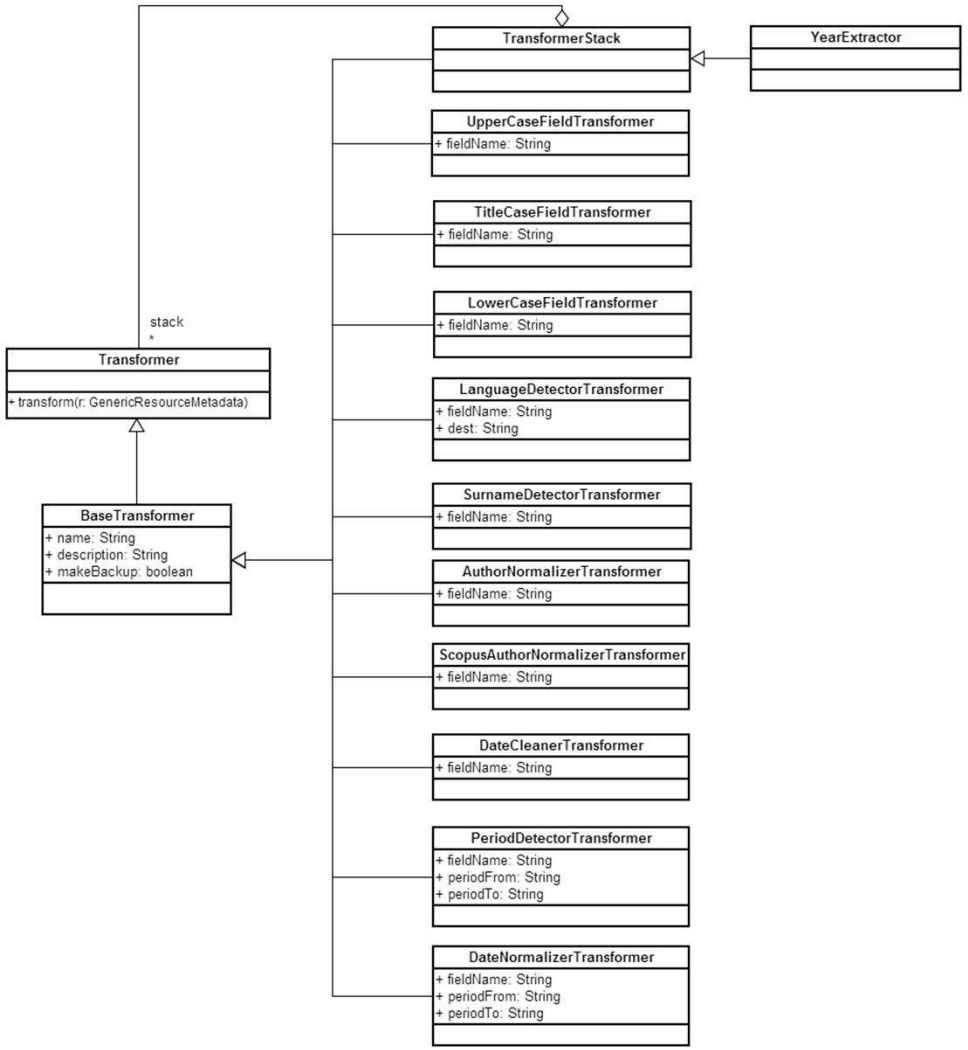

Figura 9: Filtros Desarrollados

- Aporte
- Metodología
- Filtros Desarrollados
  - **Estandarización del Formato del Texto**
    - UpperCaseFieldTransformer
    - LowerCaseFieldTransformer
    - TitleCaseFieldTransformer
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

Estandarización de Formato del Texto

- Objetivo: corregir sintácticamente el contenido de un texto a través de la aplicación de distintos formatos

- Además:
  - Eliminar espacios en blanco al inicio y al final de la cadena
  - Reducir a uno la cantidad de espacios entre palabra y palabra

- Filtros:
  - UpperCaseField
  - LowerCaseField
  - TitleCaseField

Ejemplos

```
- <metadata>
  - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
      <dc:title>Estudio de las características fisicoquímicas y organolépticas en el desarrollo de un aderezo de base vegetal con propiedades
        funcionales</dc:title>
      <dc:creator>Sosa, Carola Andrea</dc:creator>
      <dc:subject>Ingeniería</dc:subject>
      <dc:subject>Ingeniería</dc:subject>
      <dc:subject>Industria alimentaria</dc:subject>
      <dc:description>Contiene: Introducción - Introducción general - Objetivo general - Objetivos específicos - Materiales y métodos - Ensayos preliminares -
        Operaciones preparatorias - Formulación de las salsas mezcla - Aditivos - Resultados - Almacenamiento de salsas mezcla - Almacenamiento de salsas
        tratadas térmicamente - Almacenamiento de salsas preservadas con tratamientos combinados - Almacenamiento de salsas preservadas con sorbato de
        potasio - Discusiones generales - Conclusiones finales.</dc:description>
      <dc:contributor>Bevilacqua, Alicia</dc:contributor>
      <dc:contributor>Sgroppo, Sonia Cecilia</dc:contributor>
      <dc:date>2010-08-03T03:00:00Z</dc:date>
      <dc:date>2009</dc:date>
      <dc:date>2009</dc:date>
      <dc:type>Tesis</dc:type>
      <dc:type>Tesis de doctorado</dc:type>
      <dc:identifier>http://hdl.handle.net/10915/1418</dc:identifier>
      <dc:language>es</dc:language>
    </oai_dc:dc>
  </metadata>
```

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
    - **UpperCaseFieldTransformer**
    - LowerCaseFieldTransformer
    - TitleCaseFieldTransformer
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

UpperCaseFieldTransformer

```
- <record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/1418</field>
    <field name="author">Sosa, Carola Andrea</field>
    <field name="colaborator">Bevilacqua, Alicia</field>
    <field name="colaborator">Sgroppo, Sonia Cecilia</field>
    <field name="title">ESTUDIO DE LAS CARACTERÍSTICAS FISICOQUÍMICAS Y ORGANOLÉPTICAS EN EL DESARROLLO DE UN ADEREZO DE BASE VEGETAL CON
      PROPIEDADES FUNCIONALES</field>
    <field name="setSpec">hdl_10915_3</field>
    <field name="description">Contiene: Introducción - Introducción general - Objetivo general - Objetivos específicos - Materiales y métodos - Ensayos preliminares -
      Operaciones preparatorias - Formulación de las salsas mezcla - Aditivos - Resultados - Almacenamiento de salsas mezcla - Almacenamiento de salsas tratadas
      térmicamente - Almacenamiento de salsas preservadas con tratamientos combinados - Almacenamiento de salsas preservadas con sorbato de potasio -
      Discusiones generales - Conclusiones finales.</field>
    <field name="subject">Ingeniería</field>
    <field name="subject">Industria alimentaria</field>
    <field name="electronic-url">http://hdl.handle.net/10915/1418</field>
    <field name="language">es</field>
    <field name="date">2010-08-03T03:00:00Z</field>
    <field name="date">2009</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
    <field name="title_bkp">Estudio de las características fisicoquímicas y organolépticas en el desarrollo de un aderezo de base vegetal con propiedades
      funcionales</field>
  </record>
```

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
    - UpperCaseFieldTransformer
    - **LowerCaseFieldTransformer**
    - TitleCaseFieldTransformer
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

LowerCaseFieldTransformer

```
- <record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/1418</field>
    <field name="author">Sosa, Carola Andrea</field>
    <field name="colaborator">Bevilacqua, Alicia</field>
    <field name="colaborator">Sgroppo, Sonia Cecilia</field>
    <field name="title">Estudio de las características fisicoquímicas y organolépticas en el desarrollo de un aderezo de base vegetal con propiedades funcionales</field>
    <field name="setSpec">hdl_10915_3</field>
    <field name="description">contiene: introducción - introducción general - objetivo general - objetivos específicos - materiales y métodos - ensayos preliminares - operaciones preparatorias - formulación de las salsas mezcla - aditivos - resultados - almacenamiento de salsas mezcla - almacenamiento de salsas tratadas térmicamente - almacenamiento de salsas preservadas con tratamientos combinados - almacenamiento de salsas preservadas con sorbato de potasio - discusiones generales - conclusiones finales.</field>
    <field name="subject">Ingeniería</field>
    <field name="subject">Industria alimentaria</field>
    <field name="electronic-url">http://hdl.handle.net/10915/1418</field>
    <field name="description_bkp">Contiene: Introducción - Introducción general - Objetivo general - Objetivos específicos - Materiales y métodos - Ensayos preliminares - Operaciones preparatorias - Formulación de las salsas mezcla - Aditivos - Resultados - Almacenamiento de salsas mezcla - Almacenamiento de salsas tratadas térmicamente - Almacenamiento de salsas preservadas con tratamientos combinados - Almacenamiento de salsas preservadas con sorbato de potasio - Discusiones generales - Conclusiones finales.</field>
    <field name="language">es</field>
    <field name="date">2010-08-03T03:00:00Z</field>
    <field name="date">2009</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
  </record>
```

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
    - UpperCaseFieldTransformer
    - LowerCaseFieldTransformer
    - **TitleCaseFieldTransformer**
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

TitleCaseFieldTransformer

```
<record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/1418</field>
    <field name="author">Sosa, Carola Andrea</field>
    <field name="colaborator">Bevilacqua, Alicia</field>
    <field name="colaborator">Sgroppo, Sonia Cecilia</field>
    <field name="title">Estudio De Las Características Fisicoquímicas Y Organolépticas En El Desarrollo De Un Aderezo De Base Vegetal Con Propiedades
        Funcionales</field>
    <field name="setSpec">hdl_10915_3</field>
    <field name="description">Contiene: Introducción - Introducción general - Objetivo general - Objetivos específicos - Materiales y métodos - Ensayos preliminares -
        Operaciones preparatorias - Formulación de las salsas mezcla - Aditivos - Resultados - Almacenamiento de salsas mezcla - Almacenamiento de salsas tratadas
        térmicamente - Almacenamiento de salsas preservadas con tratamientos combinados - Almacenamiento de salsas preservadas con sorbato de potasio -
        Discusiones generales - Conclusiones finales.</field>
    <field name="subject">Ingeniería</field>
    <field name="subject">Industria alimentaria</field>
    <field name="electronic-url">http://hdl.handle.net/10915/1418</field>
    <field name="language">es</field>
    <field name="date">2010-08-03T03:00:00Z</field>
    <field name="date">2009</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
    <field name="title_bkp">Estudio de las características fisicoquímicas y organolépticas en el desarrollo de un aderezo de base vegetal con propiedades
        funcionales</field>
</record>
```

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - **Detección del Lenguaje del Texto**
    - LanguageDetectorTransformer
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

Detección del Lenguaje del Texto

- Objetivo: identificar el idioma de un texto cuando no contiene el metadato "language" asociado, o cuando no está especificado o tiene un valor que no corresponde a ningún idioma (ej.: “otro”).

- Filtro:
  - LanguageDetector

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
    - **LanguageDetectorTransformer**
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

LanguageDetector

- Utiliza la librería "Language Detection" de Cybozu Labs
- Calcula las probabilidades de los diferentes idiomas sobre las características de ortografía de un texto específico
- Se pueden detectar 49 lenguajes diferentes con una precisión de 99.8%

Ejemplos

Figura 10: CNEA – Distribución de Recursos por Lenguaje. Fuente: istec

LanguageDetectorTransformer

```
<?xml version="1.0"?>
- <records>
  - <record>
      <field name="id">oai:ricabib.cab.cnea.gov.ar:350</field>
      <field name="author">Bertona, Juan F.</field>
      <field name="title">Análisis neutrónico de las barras de control del reactor CAREM-25 haciendo uso del código MCNP.</field>
      <field name="electronic-format">application/pdf</field>
      <field name="setSpec">7374617475733D756E707562</field>
      <field name="setSpec">7375626A656374733D6E75635F696E67:6E75635F696E675F7265615F656E67</field>
      <field name="setSpec">7375626A656374733D6E75635F696E67:6E75635F696E675F7265615F636F6D</field>
      <field name="setSpec">74797065733D746865736973</field>
      <field name="description">El CAREM 25 es un diseño de reactor de producción eléctrica refrigerado y moderado por agua liviana, autopresurizado, integrado, de convección natural y con sistemas de seguridad pasivos. Actualmente el estado de desarrollo de la ingeniería alcanzado en el diseño hace relevante la realización de ciertos estudios que hacen a la performance. En este marco es de importancia la evaluación de los efectos de al radiación sobre las barras de control de Ag-In-Cd. En el presente trabajo se realizó un modelado del núcleo con el código de transporte probabilístico MCNP junto con sus barras de control y otros componentes pertinentes para el cálculo, a partir de un modelo de fuente neutrónica fija partiendo de los resultados de la cadena de cálculo CONDOR->CITVAP/THERMIT. Se obtuvieron resultados relacionados con el flujo neutrónico, el calentamiento instantáneo o prompt, el calentamiento por decaimientos nucleares, la activación y el quemado o "depletion" de los materiales de las barras de control. Entre los resultados se encontró una fuerte depletion del isótopo $^{113}$Cd que tiene como consecuencia una notable disminución en la sección eficaz macroscópica del material Ag-In-Cd.</field>
      <field name="subject">Ingeniería de reactores</field>
      <field name="subject">Componentes y consideraciones de diseño de reactores</field>
      <field name="subject">Control elements</field>
      <field name="subject">Elementos de control</field>
      <field name="subject">Heating</field>
      <field name="subject">Calentamiento</field>
      <field name="subject">Activation</field>
      <field name="subject">Activación</field>
      <field name="subject">CAREM-25</field>
      <field name="electronic-url">http://ricabib.cab.cnea.gov.ar/350/2/1Bertona_J..pdf</field>
      <field name="electronic-url">http://ricabib.cab.cnea.gov.ar/350/3/1Bertona.pdf</field>
      <field name="electronic-url">Bertona, Juan F. (2012) Análisis neutrónico de las barras de control del reactor CAREM-25 haciendo uso del código MCNP. / Neutronic [illegible] reactor´s control rods using MCNP. Proyecto Integrador Ingeniería Nuclear, Universidad Nacional de Cuyo, Instituto Balseiro.</field>
      <field name="relation">http://ricab[illegible]cab.cnea.gov.ar/350/</field>
      <field name="language">es</field>
      <field name="date">2012-06-18</field>
      <field name="[illegible]">[illegible]</field>
      <field name="medium">NonPeerReviewed</field>
    </record>
```

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
  - **Normalización de Nombre de Autor**
    - SurnameDetector
    - AuthorNormalizer
    - ScopusAuthorNormalizer
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

Normalización de Nombre de Autor

- Objetivo: Lograr que la firma de un autor aparezca citada de la misma manera en todas sus ocurrencias.

- Filtros:
  - SurnameDetector
  - AuthorNormalizer
  - ScopusAuthorNormalizer

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
    - **SurnameDetector**
    - AuthorNormalizer
    - ScopusAuthorNormalizer
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

SurnameDetector

- Detecta los apellidos de un autor, según su ocurrencia en la base de autores de SCOPUS o según su sintaxis

1. Nombre separado por comas:

   Palabra1, Palabra2 Palabra3 → Palabra1 es el apellido

2. Nombre separado por espacios:

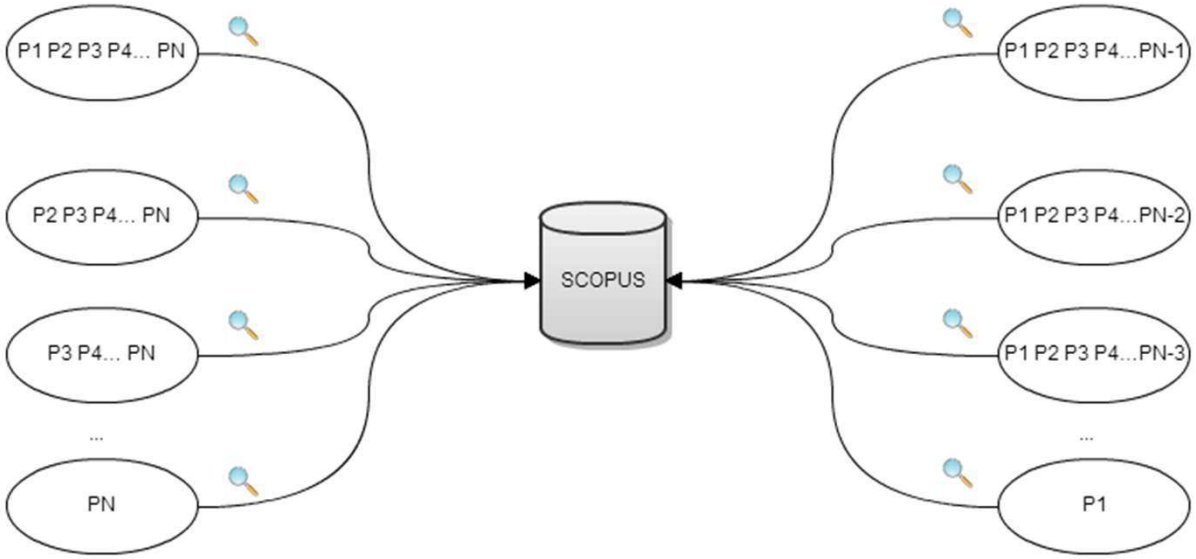

Figura 11: SurnameDetectorTransformer

Ejemplos – Caso 1

```
- <record>
  - <header>
      <identifier>oai:sedici.unlp.edu.ar:10915/1423</identifier>
      <datestamp>2012-10-18T18:30:10Z</datestamp>
      <setSpec>hdl_10915_3</setSpec>
    </header>
  - <metadata>
    - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
        <dc:title>Estudio teórico y validación experimental de la dehidrocongelación de ananá</dc:title>
        <dc:creator>Ramallo, Laura Ana</dc:creator>
        <dc:subject>Ingeniería</dc:subject>
        <dc:subject>Ingeniería química</dc:subject>
        <dc:subject>Industria alimentaria</dc:subject>
        <dc:description>Contiene: Introducción - Objetivos - Antecedentes - Secado - Deshidratación osmótica - Deshidratación osmótica con pulsos de vacío -
          Dehidrocongelación - Conclusiones generales.</dc:description>
        <dc:contributor>Mascheroni, Rodolfo Horacio</dc:contributor>
        <dc:date>2010-08-03T03:00:00Z</dc:date>
        <dc:date>2010</dc:date>
        <dc:date>2010</dc:date>
        <dc:type>Tesis</dc:type>
        <dc:type>Tesis de doctorado</dc:type>
        <dc:identifier>http://hdl.handle.net/10915/1423</dc:identifier>
        <dc:language>es</dc:language>
      </oai_dc:dc>
    </metadata>
  </record>
```

Resultado

```
<record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/1423</field>
    <field name="author">Ramallo, Laura Ana</field>
    <field name="colaborator">Mascheroni, Rodolfo Horacio</field>
    <field name="title">Estudio teórico y validación experimental de la dehidrocongelación de ananá</field>
    <field name="setSpec">hdl_10915_3</field>
    <field name="description">Contiene: Introducción - Objetivos - Antecedentes - Secado - Deshidratación osmótica - Deshidratación osmótica con pulsos de vacío -
        Dehidrocongelación - Conclusiones generales.</field>
    <field name="subject">Ingeniería</field>
    <field name="subject">Industria alimentaria</field>
    <field name="electronic-url">http://hdl.handle.net/10915/1423</field>
    <field name="language">es</field>
    <field name="date">2010-08-03T03:00:00Z</field>
    <field name="date">2010</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
</record>
```

Ejemplos – Caso 2

```
<record>
  - <header>
      <identifier>oai:sedici.unlp.edu.ar:10915/1431</identifier>
      <datestamp>2012-10-18T18:48:04Z</datestamp>
      <setSpec>hdl_10915_3</setSpec>
    </header>
  - <metadata>
    - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
        <dc:title>Mesoestructura, comportamiento mecánico y propiedades de transporte en hormigón</dc:title>
        <dc:creator>María Celeste Torrijos</dc:creator>
        <dc:subject>Ingeniería</dc:subject>
        <dc:subject>Hormigón</dc:subject>
        <dc:subject>Ingeniería</dc:subject>
        <dc:subject>Tecnología de la construcción</dc:subject>
        <dc:description>Contiene: Introducción - Estructura y propiedades del hormigón - Programa experimental - Estudio sobre hormigones dañados - Estudios
          sobre hormigones reforzados con fibras y hormigones autocompactantes - Conclusiones - Anexos.</dc:description>
        <dc:contributor>Zerbino, Raúl Luis</dc:contributor>
        <dc:date>2010-08-03T03:00:00Z</dc:date>
        <dc:date>2008</dc:date>
        <dc:date>2008</dc:date>
        <dc:type>Tesis</dc:type>
        <dc:type>Tesis de doctorado</dc:type>
        <dc:identifier>http://hdl.handle.net/10915/1431</dc:identifier>
        <dc:language>es</dc:language>
      </oai_dc:dc>
    </metadata>
</record>
```

SCOPUS

```
<doc>
     <field name="affiliation"><![CDATA[Torrijos, M.C., CONICET-LEMIT, Fac. Ing. UNLP, 52 entre 121 y 122, La Plata, 1900, Argentina]]></field>
     <field name="affiliation"><![CDATA[Barragán, B.E., UPC, Jordi Girona 1-3 M.C1, Barcelona, 08034, Spain]]></field>
     <field name="affiliation"><![CDATA[Zerbino, R.L., CONICET-LEMIT, Fac. Ing. UNLP, 52 entre 121 y 122, La Plata, 1900, Argentina]]></field>
     <field name="year">2008</field>
</doc>
<doc>
     <field name="affiliation"><![CDATA[Vezzani, D., Ecología de Reservorios y Vectores de Parásitos, Facultad de Cs. Exactas y Naturales, Universidad de Buenos
     Aires, Argentina]]></field>
     <field name="affiliation"><![CDATA[Fontanarrosa, M.F., Laboratorio DIAP (Diagnóstico en Animales Pequeños), Inca 109, (B1836BBC), Llavallol, Buenos Aires,
     Argentina]]></field>
     <field name="affiliation"><![CDATA[Eiras, D.F., Laboratorio DIAP (Diagnóstico en Animales Pequeños), Inca 109, (B1836BBC), Llavallol, Buenos Aires,
     Argentina, Cátedra de Parasitología y Enfermedades Parasitarias, Departamento de Microbiología, Facultad de Ciencias Veterinarias, CC 296, B1900AVW La Plata,
     Argentina]]></field>
     <field name="year">2008</field>
</doc>
```

Resultado

```
<record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/1431</field>
    <field name="author">Torrijos, María Celeste</field>
    <field name="colaborator">Zerbino, Raúl Luis</field>
    <field name="title">Mesoestructura, comportamiento mecánico y propiedades de transporte en hormigón</field>
    <field name="setSpec">hdl_10915_3</field>
    <field name="description">Contiene: Introducción - Estructura y propiedades del hormigón - Programa experimental - Estudio sobre hormigones dañados - Estudios
        sobre hormigones reforzados con fibras y hormigones autocompactantes - Conclusiones - Anexos.</field>
    <field name="subject">Ingeniería</field>
    <field name="subject">Hormigón</field>
    <field name="subject">Tecnología de la construcción</field>
    <field name="electronic-url">http://hdl.handle.net/10915/1431</field>
    <field name="language">es</field>
    <field name="date">2010-08-03T03:00:00Z</field>
    <field name="date">2008</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
</record>
```

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
    - SurnameDetector
    - **AuthorNormalizer**
    - ScopusAuthorNormalizer
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

AuthorNormalizer

1. Utiliza SurnameDetector para detectar el apellido de un autor.

2. Normaliza el nombre dejando sólo sus iniciales:

   Apellidos, Iniciales

Ejemplo

```
- <record>
  - <header>
      <identifier>oai:sedici.unlp.edu.ar:10915/4508</identifier>
      <datestamp>2012-10-24T18:43:25Z</datestamp>
      <setSpec>hdl_10915_42</setSpec>
    </header>
  - <metadata>
    - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
        <dc:title>Estudio de la producción de lacasas fúngicas extracelulares en diferentes cepas autóctonas</dc:title>
        <dc:creator>Saparrat, Mario Carlos Nazareno</dc:creator>
        <dc:subject>Ciencias Naturales</dc:subject>
        <dc:subject>Botánica</dc:subject>
        <dc:subject>Hongos</dc:subject>
        <dc:subject>Enzimas</dc:subject>
        <dc:description>Se estudió la producción de lacasas y otras oxidorreductasas extracelulares relacionadas con los sistemas ligninolíticos fúngicos, en
          diferentes cepas autóctonas. Las lacasas son enzimas con amplia especificidad de sustrato, involucradas en la pudrición blanca de la madera, en otros
          aspectos de la fisiología fúngica, como pigmentación conidial, procesos de morfogénesis y patogénesis, y en la detoxificación de compuestos fenólicos
          derivados de la ligninólisis, y fitoalexinas. La selección de cepas productoras, aisladas a partir de diferentes sustratos, es la primera aproximación en
          la detección de los sistemas enzimáticos oxidativos extracelulares. Se detectó actividad oxidorreductasa extracelular, utilizando tests en placa
          conteniendo el cromóforo ABTS: ácido 2,2'-azino-bis (3-etilbenzotialina-6-sulfónico), en diferentes cepas de Ascomycetes, Basidiomycetes y
          Deiíleromycetes; se reportan por primera vez, diferentes especies con actividad enzimática oxidativa extracelular.</dc:description>
        <dc:contributor>Arambarri, Angélica Margarita</dc:contributor>
        <dc:contributor>Cabello, Marta Noemí</dc:contributor>
        <dc:date>2010-10-22T03:00:00Z</dc:date>
        <dc:date>2000</dc:date>
        <dc:date>2000</dc:date>
        <dc:type>Tesis</dc:type>
        <dc:type>Tesis de doctorado</dc:type>
        <dc:identifier>http://hdl.handle.net/10915/4508</dc:identifier>
        <dc:language>es</dc:language>
      </oai_dc:dc>
    </metadata>
  </record>
```

Resultado

```
<record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/4508</field>
    <field name="author">Saparrat, M.C.N.</field>
    <field name="colaborator">Arambarri, Angélica Margarita</field>
    <field name="colaborator">Cabello, Marta Noemí</field>
    <field name="title">Estudio de la producción de lacasas fúngicas extracelulares en diferentes cepas autóctonas</field>
    <field name="setSpec">hdl_10915_42</field>
    <field name="description">Se estudió la producción de lacasas y otras oxidorreductasas extracelulares relacionadas con los sistemas ligninolíticos fúngicos, en
        diferentes cepas autóctonas. Las lacasas son enzimas con amplia especificidad de sustrato, involucradas en la pudrición blanca de la madera, en otros aspectos
        de la fisiología fúngica, como pigmentación conidial, procesos de morfogénesis y patogénesis, y en la detoxificación de compuestos fenólicos derivados de la
        ligninólisis, y fitoalexinas. La selección de cepas productoras, aisladas a partir de diferentes sustratos, es la primera aproximación en la detección de los
        sistemas enzimáticos oxidativos extracelulares. Se detectó actividad oxidorreductasa extracelular, utilizando tests en placa conteniendo el cromóforo ABTS:
        ácido 2,2'-azino-bis (3-etilbenzotialina-6-sulfónico), en diferentes cepas de Ascomycetes, Basidiomycetes y Deiíleromycetes; se reportan por primera vez,
        diferentes especies con actividad enzimática oxidativa extracelular.</field>
    <field name="subject">Ciencias Naturales</field>
    <field name="subject">Botánica</field>
    <field name="subject">Hongos</field>
    <field name="subject">Enzimas</field>
    <field name="electronic-url">http://hdl.handle.net/10915/4508</field>
    <field name="language">es</field>
    <field name="date">2010-10-22T03:00:00Z</field>
    <field name="date">2000</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
</record>
```

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
    - SurnameDetector
    - AuthorNormalizer
    - **ScopusAuthorNormalizer**
  - Fecha: Depuración y Normalización
- Conclusiones
- Trabajos Futuros

ScopusAuthorNormalizer

1. Utiliza SurnameDetector para detectar el apellido de un autor.

2. Utiliza AuthorNormalizer para normalizar el nombre a la forma *Apellidos, Iniciales*

3. Normaliza el nombre del autor según su ocurrencia en la base de datos de SCOPUS dependiendo el caso

...ScopusAuthorNormalizer

Entrada	Aparición en Scopus (Apellido, A. B.*)	Salida (Apellidos, A. C.*)	Nivel Confianza (Bajo, Medio, Alto)
Apellidos, A. B.	-	Apellidos, A. B.	Alto
Apellidos, A.	Apellidos, A. B.	Apellidos, A.	Alto
Apellidos, A.	Apellidos, B. A.	Apellidos, A.	Alto
Apellidos, A. B.	Apellidos, A.	Apellidos, A.	Alto
Apellidos, A. B.	Apellidos, B. A.	Apellidos, B. A.	Medio
Apellidos, A. B.	Apellidos, A. B.	Apellidos, A. B.	Alto
Apellidos, A. B.	Apellidos, C.	Apellidos, A. B.	Alto

Tabla 1: ScopusAuthorNormalizerTransformer – Casos Posibles

Ejemplo

```
<metadata>
  - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
        <dc:title>Parques eólicos con conexión a redes debiles</dc:title>
        <dc:creator>Fernández, Roberto Daniel</dc:creator>
        <dc:subject>Ingeniería</dc:subject>
        <dc:subject>Energia</dc:subject>
        <dc:subject>Energía eólica</dc:subject>
        <dc:description>Dado que este trabajo de tesis busca, por medio del análisis, entendimiento y operación de las granjas eólicas, poder plantear criterios de
            control que puedan contribuir a la estabilidad de la red eléctrica de la cual forman parte, se evalúan criterios para las granjas eólicas operadas como
            centrales generadoras buscando, a la vez, explotar al máximo las características del aprovechamiento. Los principales aportes presentados en esta
            tesis se resumen en los siguientes puntos: 1. Proponer leyes de control para las granjas de manera de asegurar que den soporte a la red eléctrica
            tanto desde el punto de vista de la frecuencia como de la tensión. 2. Proponer leyes de control para las granjas de manera de asegurar que ellas
            contribuyan a la estabilidad de la red a la cual se encuentren conectadas. 3. Respecto de la potencia activa de las granjas se proponen leyes de control
            que implican un comportamiento similar al de los generadores sincrónicos de los modernos sistemas de generación convencional (Cap. 6). 4. Respecto
            de la potencia reactiva de las granjas se propone una ley de control que permite que las granjas contribuyan al perfil de la tensión en el punto de
            conexión (Cap. 6). 5. Incorporar al formalismo matemático de las redes eléctricas convencionales, la propuesta, estudio, análisis e impacto de
            distintas granjas eólicas según las máquinas eléctricas que las componen y teniendo en cuenta el porcentaje de penetración eólica respecto de la
            potencia generada por medios convencionales (Cap. 7). 6. Incorporar dentro del formalismo matemático precedente las leyes de control lineales tanto
            para las potencias activa como reactiva de manera de verificar analíticamente el impacto de las propuestas (Cap. 7). 7. Proponer lazos de control no
            lineales tanto para las potencias activa como reactiva de las granjas de manera de asegurar la contribución de las mismas a la estabilidad de la red
            eléctrica (Cap. 8). 8. Asegurar, en todos los casos, la practicidad de las propuestas de control presentadas de manera que la implementación de las
            mismas sea sencilla.</dc:description>
        <dc:contributor>Mantz, Ricardo J.</dc:contributor>
        <dc:contributor>Battaiotto, Pedro Eduardo</dc:contributor>
        <dc:date>2010-08-03T03:00:00Z</dc:date>
        <dc:date>2007</dc:date>
        <dc:date>2007</dc:date>
        <dc:type>Tesis</dc:type>
        <dc:type>Tesis de doctorado</dc:type>
        <dc:identifier>http://hdl.handle.net/10915/1435</dc:identifier>
        <dc:language>es</dc:language>
    </oai_dc:dc>
</metadata>
```

SCOPUS

```
<doc>
     <field name="affiliation"><![CDATA[Argerami, M., Department of Mathematics, University of Regina, Regina, SK, Canada]]></field>
     <field name="affiliation"><![CDATA[Massey, P., Departamento de Matemática, Universidad Nacional de la Plata, Instituto Argentino de Matemática-conicet,
     Argentina]]></field>
     <field name="year">2009</field>
</doc>
<doc>
     <field name="affiliation"><![CDATA[de Brodtkorb, M.K., CONICET-Universidad de Buenos Aires, Argentina]]></field>
     <field name="affiliation"><![CDATA[Pezzutti, N., Geóloga Consultora, Argentina]]></field>
     <field name="affiliation"><![CDATA[Poma, S., CONICET-Universidad de Buenos Aires, Argentina]]></field>
     <field name="affiliation"><![CDATA[Fernández, R., Instituto de Recursos Minerales, FCNyM-CICBA, Universidad Nacional de La Plata, Argentina]]></field>
     <field name="year">2009</field>
</doc>
```

Resultado

```
<record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/1435</field>
    <field name="author">Fernández, R.</field>
    <field name="colaborator">Mantz, Ricardo J.</field>
    <field name="colaborator">Battaiotto, Pedro Eduardo</field>
    <field name="title">Parques eólicos con conexión a redes débiles</field>
    <field name="setSpec">hdl_10915_3</field>
    <field name="description">Dado que este trabajo de tesis busca, por medio del análisis, entendimiento y operación de las granjas eólicas, poder plantear criterios de
        control que puedan contribuir a la estabilidad de la red eléctrica de la cual forman parte, se evalúan criterios para las granjas eólicas operadas como centrales
        generadoras buscando, a la vez, explotar al máximo las características del aprovechamiento. Los principales aportes presentados en esta tesis se resumen en los
        siguientes puntos: 1. Proponer leyes de control para las granjas de manera de asegurar que den soporte a la red eléctrica tanto desde el punto de vista de la
        frecuencia como de la tensión. 2. Proponer leyes de control para las granjas de manera de asegurar que ellas contribuyan a la estabilidad de la red a la cual se
        encuentren conectadas. 3. Respecto de la potencia activa de las granjas se proponen leyes de control que implican un comportamiento similar al de los
        generadores sincrónicos de los modernos sistemas de generación convencional (Cap. 6). 4. Respecto de la potencia reactiva de las granjas se propone una ley de
        control que permite que las granjas contribuyan al perfil de la tensión en el punto de conexión (Cap. 6). 5. Incorporar al formalismo matemático de las redes
        eléctricas convencionales, la propuesta, estudio, análisis e impacto de distintas granjas eólicas según las máquinas eléctricas que las componen y teniendo en
        cuenta el porcentaje de penetración eólica respecto de la potencia generada por medios convencionales (Cap. 7). 6. Incorporar dentro del formalismo
        matemático precedente las leyes de control lineales tanto para las potencias activa como reactiva de manera de verificar analíticamente el impacto de las
        propuestas (Cap. 7). 7. Proponer lazos de control no lineales tanto para las potencias activa como reactiva de las granjas de manera de asegurar la contribución
        de las mismas a la estabilidad de la red eléctrica (Cap. 8). 8. Asegurar, en todos los casos, la practicidad de las propuestas de control presentadas de manera que
        la implementación de las mismas sea sencilla.</field>
    <field name="subject">Ingeniería</field>
    <field name="subject">Energia</field>
    <field name="subject">Energía eólica</field>
    <field name="electronic-url">http://hdl.handle.net/10915/1435</field>
    <field name="language">es</field>
    <field name="date">2010-08-03T03:00:00Z</field>
    <field name="date">2007</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
</record>
```

Normalización de Nombre de Autor

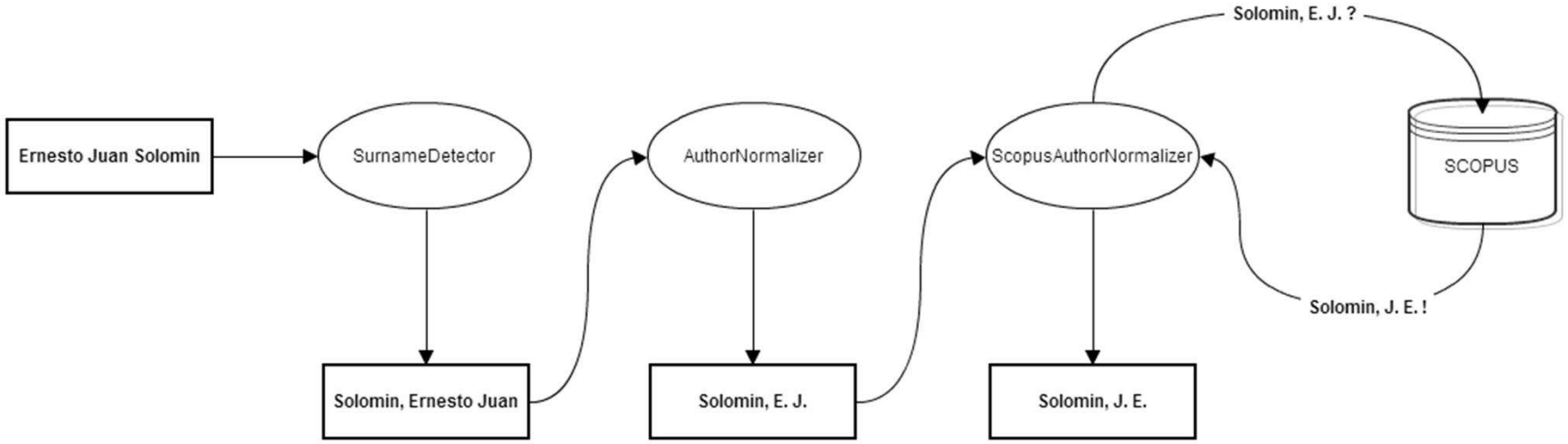

Figura 12: Normalización de Nombre de Autor – Evolución de Filtros

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
  - **Fecha: Depuración y Normalización**
    - DateCleaner
    - PeriodDEtector
    - DateNormalizer
- Conclusiones
- Trabajos Futuros

Fecha: Depuración y Normalización

- Objetivo: estandarizar el formato del metadato "date", según la norma ISO 8601 → "yyyy-MM-dd"

- Contemplar la mayoría de las diferentes notaciones de fechas
  - Little-endian gregoriano
  - Big-endian gregoriano

- Filtros:
  - DateCleaner
  - PeriodDetector
  - DateNormalizer

Little-Endian / Big-Endian

Little-Endian	Big-endian
8 November 2003, 8. November 2003	2003 November 9
8/11/2003, 08.11.2003, 8-11-2003	2003.11.9, 2003-11-09, 2003. 11. 9.
08-Nov-2003 , 08Nov03	2003Nov9, 2003Nov09, 2003-Nov-9, 2003-Nov-09
[The] 8th [of] November 2003	2003. november 9., 2003. nov. 9.
Sunday, 8 November 2003	2003-Nov-9, Sunday
8/xi/03, 8.xi.03, 8-xi.03, o 8.XI.2003	2003. XI. 9.
8 November AD 2003	9 November 2003, 18h 14m 12s , 2003/11/9/18:14:12, 2003-11-09T18:14:12

Tabla 2: Formatos de fecha contemplados - Ejemplos

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
    - **DateCleaner**
    - PeriodDetector
    - DateNormalizer
- Conclusiones
- Trabajos Futuros

DateCleaner

- Elimina caracteres basura, como‘[]’, blancos repetidos
- Convierte signos como ‘?’, ‘/’ o ‘.’ a ‘-’
- Elimina todo tipo de construcciones inválidas o aclaraciones, por ejemplo ‘(4° fecha)’ .

Ejemplo

```
<record>
  - <header>
        <identifier>oai:sedici.unlp.edu.ar:10915/1424</identifier>
        <datestamp>2012-10-18T18:31:12Z</datestamp>
        <setSpec>hdl_10915_3</setSpec>
    </header>
  - <metadata>
      - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
            <dc:title>Estudio teórico-experimental de los fenómenos de adsorción relacionados con la reacción del electrodo de hidrógeno</dc:title>
            <dc:creator>Chialvo, Abel César</dc:creator>
            <dc:subject>Ingeniería</dc:subject>
            <dc:subject>Cinética química</dc:subject>
            <dc:subject>Ingeniería</dc:subject>
            <dc:description>Contiene: Antecedentes de la reacción del electrodo de hidrógeno (HER) - Descripción clásica de la cinética de la HER - Tratamiento
                cinético generalizado de la HER - Evaluación de los parámetros cinéticos elementales de la HER - El intermediario de reacción en la HER - Análisis y
                discusión - Conclusiones.</dc:description>
            <dc:contributor>Triaca, W. E.</dc:contributor>
            <dc:contributor>Vicintin, Arnaldo</dc:contributor>
            <dc:date>2010/08/03</dc:date>
            <dc:date>2007</dc:date>
            <dc:type>Tesis</dc:type>
            <dc:type>Tesis de doctorado</dc:type>
            <dc:identifier>http://hdl.handle.net/10915/1424</dc:identifier>
            <dc:language>es</dc:language>
        </oai_dc:dc>
    </metadata>
</record>
```

Resultado

```
<record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/1424</field>
    <field name="author">Chialvo, Abel César</field>
    <field name="colaborator">Triaca, W. E.</field>
    <field name="colaborator">Visintin, Arnaldo</field>
    <field name="title">Estudio teórico-experimental de los fenómenos de adsorción relacionados con la reacción del electrodo de hidrógeno</field>
    <field name="setSpec">hdl_10915_3</field>
    <field name="description">Contiene: Antecedentes de la reacción del electrodo de hidrógeno (HER) - Descripción clásica de la cinética de la HER - Tratamiento
        cinético generalizado de la HER - Evaluación de los parámetros cinéticos elementales de la HER - El intermediario de reacción en la HER - Análisis y discusión -
        Conclusiones.</field>
    <field name="subject">Ingeniería</field>
    <field name="subject">Cinética química</field>
    <field name="electronic-url">http://hdl.handle.net/10915/1424</field>
    <field name="language">es</field>
    <field name="date">2010-08-03</field>
    <field name="date">2007</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
</record>
```

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
    - DateCleaner
    - **PeriodDetector**
    - DateNormalizer
- Conclusiones
- Trabajos Futuros

PeriodDetector

1. Limpia el dato con el filtro “DateCleaner”

2. Revisa el dato de entrada intentando detectar períodos de acuerdo a determinados formatos (ej.: 'dd' al 'dd' de MMMM de yyyy, yyyy-yyyy, from yyyy to yyyy, etc.).

3. Si encuentra un período, separa el campo en dos, una fecha inicial y una final (dateFrom, dateTo).

Ejemplo

```
<metadata>
  - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
        <dc:title>Histofisiología del páncreas endocrino de la rata en un modelo de envejecimiento normal y un modelo diabético espontáneo</dc:title>
        <dc:creator>Riccillo, Fernando L.</dc:creator>
        <dc:subject>Ciencias Naturales</dc:subject>
        <dc:subject>Biología</dc:subject>
        <dc:subject>Fisiología animal</dc:subject>
        <dc:subject>Endocrinología</dc:subject>
        <dc:subject>Roedores</dc:subject>
        <dc:description>El objetivo de nuestro trabajo fue comparar la evolución del páncreas endocrino durante la vida de dos cepas murinas, una normal
            (Sprague Dawley) y una diabética espontánea (eSMT) a través del estudio de diferentes variables morfológicas y plasmáticas, con el propósito de
            analizar las alteraciones observadas en dichas líneas y aportar nueva información referente al desarrollo de la patología diabética tipo 2. En el
            presente estudio se utilizaron ratas macho de 4, 24 y 30 meses de edad de una cepa normal (Sprague Dawley); y de 2, 5, 10, 14 y 18 meses de una
            línea diabética espontánea (eSMT). Los individuos fueron mantenidos a 22 ± 2ºC, bajo un ciclo de 12/12 horas luz / oscuridad con una dieta estándar
            y agua ad libitum. Los páncreas extraídos fueron fijados en líquido de Bouin para su posterior procesamiento para microscopía de luz (HyE, tricrómico
            de Gomori e inmunohistoquímica). Para análisis específicos, los cortes histológicos se inmunomarcaron para la detección de insulina, glucagón,
            somatostatina, PP y antígeno nuclear de proliferación celular (PCNA), mediante el empleo de anticuerpos primarios comerciales y anticuerpos
            secundarios conjugados con peroxidasa o fosfatasa alcalina. Para los estudios morfométricos se utilizó un sistema de videomicroscopía digital
            (software Optimas®) para el cálculo de la densidad de volumen (DV) de las diferentes poblaciones endocrinas. Los niveles plasmáticos de glucosa,
            triglicéridos y colesterol se analizaron mediante kits comerciales, mientras que la insulina se midió por radioinmunoensayo (RIA).</dc:description>
        <dc:contributor>Cónsole, Gloria Miriam</dc:contributor>
        <dc:contributor>Ronderos, Jorge Rafael</dc:contributor>
        <dc:date>2010/2011</dc:date>
        <dc:date>2007</dc:date>
        <dc:type>Tesis</dc:type>
        <dc:type>Tesis de doctorado</dc:type>
        <dc:identifier>http://hdl.handle.net/10915/4443</dc:identifier>
        <dc:language>es</dc:language>
    </oai_dc:dc>
</metadata>
```

Resultado

```
<record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/4443</field>
    <field name="author">Riccillo, Fernando L.</field>
    <field name="colaborator">Cónsole, Gloria Miriam</field>
    <field name="colaborator">Ronderos, Jorge Rafael</field>
    <field name="title">Histofisiología del páncreas endocrino de la rata en un modelo de envejecimiento normal y un modelo diabético espontáneo</field>
    <field name="setSpec">hdl_10915_42</field>
    <field name="description">El objetivo de nuestro trabajo fue comparar la evolución del páncreas endocrino durante la vida de dos cepas murinas, una normal
        (Sprague Dawley) y una diabética espontánea (eSMT) a través del estudio de diferentes variables morfológicas y plasmáticas, con el propósito de analizar las
        alteraciones observadas en dichas líneas y aportar nueva información referente al desarrollo de la patología diabética tipo 2. En el presente estudio se utilizaron
        ratas macho de 4, 24 y 30 meses de edad de una cepa normal (Sprague Dawley); y de 2, 5, 10, 14 y 18 meses de una línea diabética espontánea (eSMT). Los
        individuos fueron mantenidos a 22 ± 2ºC, bajo un ciclo de 12/12 horas luz / oscuridad con una dieta estándar y agua ad libitum. Los páncreas extraídos fueron
        fijados en líquido de Bouin para su posterior procesamiento para microscopía de luz (HyE, tricrómico de Gomori e inmunohistoquímica). Para análisis específicos,
        los cortes histológicos se inmunomarcaron para la detección de insulina, glucagón, somatostatina, PP y antígeno nuclear de proliferación celular (PCNA),
        mediante el empleo de anticuerpos primarios comerciales y anticuerpos secundarios conjugados con peroxidasa o fosfatasa alcalina. Para los estudios
        morfométricos se utilizó un sistema de videomicroscopía digital (software Optimas®) para el cálculo de la densidad de volumen (DV) de las diferentes
        poblaciones endocrinas. Los niveles plasmáticos de glucosa, triglicéridos y colesterol se analizaron mediante kits comerciales, mientras que la insulina se midió
        por radioinmunoensayo (RIA).</field>
    <field name="subject">Ciencias Naturales</field>
    <field name="subject">Biología</field>
    <field name="subject">Fisiología animal</field>
    <field name="subject">Endocrinología</field>
    <field name="subject">Roedores</field>
    <field name="electronic-url">http://hdl.handle.net/10915/4443</field>
    <field name="language">es</field>
    <field name="dateFrom">2010</field>
    <field name="dateTo">2011</field>
    <field name="date">2007</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
</record>
```

- Aporte
- Metodología
- Filtros Desarrollados
  - Estandarización del Formato del Texto
  - Detección del Lenguaje del Texto
  - Normalización de Nombre de Autor
  - Fecha: Depuración y Normalización
    - DateCleaner
    - PeriodDEtector
    - **DateNormalizer**
- Conclusiones
- Trabajos Futuros

DateNormalizer

1. Emplea el filtro “DateCleaner” para limpiar valores indeseados

2. Utiliza “PeriodDetector”, para detectar períodos en el caso que hubiese

3. Normaliza tanto el campo date, como dateFrom y dateTo para los documentos que posean un período como fecha, al formato “yyyy-MM-dd”.

Ejemplo

```
<metadata>
  - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
        <dc:title>Estudio de las características fisicoquímicas y organolépticas en el desarrollo de un aderezo de base vegetal con propiedades
            funcionales</dc:title>
        <dc:creator>Sosa, Carola Andrea</dc:creator>
        <dc:subject>Ingeniería</dc:subject>
        <dc:subject>Ingeniería</dc:subject>
        <dc:subject>Industria alimentaria</dc:subject>
        <dc:description>Contiene: Introducción - Introducción general - Objetivo general - Objetivos específicos - Materiales y métodos - Ensayos preliminares -
            Operaciones preparatorias - Formulación de las salsas mezcla - Aditivos - Resultados - Almacenamiento de salsas mezcla - Almacenamiento de salsas
            tratadas térmicamente - Almacenamiento de salsas preservadas con tratamientos combinados - Almacenamiento de salsas preservadas con sorbato de
            potasio - Discusiones generales - Conclusiones finales.</dc:description>
        <dc:contributor>Bevilacqua, Alicia</dc:contributor>
        <dc:contributor>Sgroppo, Sonia Cecilia</dc:contributor>
        <dc:date>2010-08-03T03:00:00Z</dc:date>
        <dc:date>2009</dc:date>
        <dc:type>Tesis</dc:type>
        <dc:type>Tesis de doctorado</dc:type>
        <dc:identifier>http://hdl.handle.net/10915/1418</dc:identifier>
        <dc:language>es</dc:language>
    </oai_dc:dc>
</metadata>
```

Resultado

```
<record>
    <field name="id">oai:sedici.unlp.edu.ar:10915/1418</field>
    <field name="author">Sosa, Carola Andrea</field>
    <field name="colaborator">Bevilacqua, Alicia</field>
    <field name="colaborator">Sgroppo, Sonia Cecilia</field>
    <field name="title">Estudio de las características fisicoquímicas y organolépticas en el desarrollo de un aderezo de base vegetal con propiedades funcionales</field>
    <field name="setSpec">hdl_10915_3</field>
    <field name="description">Contiene: Introducción - Introducción general - Objetivo general - Objetivos específicos - Materiales y métodos - Ensayos preliminares -
        Operaciones preparatorias - Formulación de las salsas mezcla - Aditivos - Resultados - Almacenamiento de salsas mezcla - Almacenamiento de salsas tratadas
        térmicamente - Almacenamiento de salsas preservadas con tratamientos combinados - Almacenamiento de salsas preservadas con sorbato de potasio -
        Discusiones generales - Conclusiones finales.</field>
    <field name="subject">Ingeniería</field>
    <field name="subject">Industria alimentaria</field>
    <field name="electronic-url">http://hdl.handle.net/10915/1418</field>
    <field name="language">es</field>
    <field name="date">2009</field>
    <field name="date">2010-08-03</field>
    <field name="medium">Tesis</field>
    <field name="medium">Tesis de doctorado</field>
</record>
```

...Fecha: Depuración y Normalización

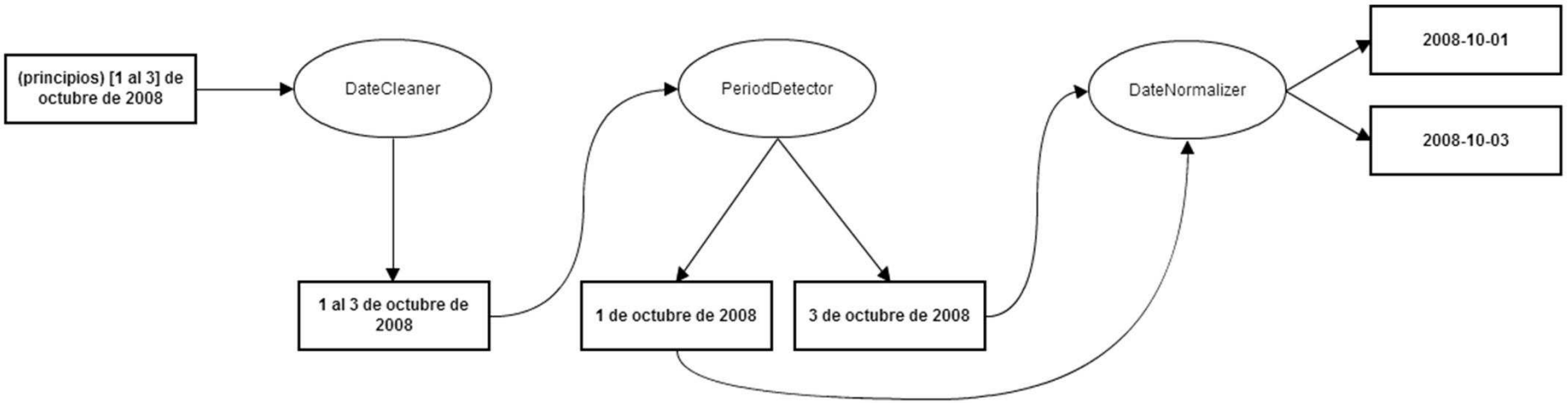

Figura 13: Depuración y Normalización de Fecha– Evolución de Filtros

- Aporte
- Metodología
- Filtros Desarrollados
- **Conclusiones**
- Trabajos Futuros

- Los RI constituyen una fuente de datos para estudios y estadísticas vinculadas a la producción científica de una institución

- Es necesario que los datos estén normalizados para poder explotar al máximo la información que contiene

Conclusiones

- ✓ Se desarrollaron métodos que mejoran la calidad de los datos contenidos en SeDiCI:
  - ✓ Se optimiza el uso y se maximiza el aprovechamiento del material
  - ✓ Se facilitan los procesos de recuperación de información
  - ✓ Se optimizan los procesos de intercambio de información

- Datos que garantizan cierto nivel de calidad, aseguran una mejor exposición de la producción científica de la institución.

- Aporte
- Metodología
- Filtros Desarrollados
- Conclusiones
- **Trabajos Futuros**

Trabajos Futuros

- Optimización de los procesos de análisis y transformación a fin de mejorar su performance
- Normalización/Incorporación de la afiliación de un autor
- Normalización del lugar de publicación de un documento
- Integración con tecnologías semánticas

¡Muchas Gracias!

¿Preguntas?

Almazán, María Belén

belenalmazan@gmail.com

SeDiCI
SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL

UNLP

FACULTAD DE INFORMÁTICA | UNLP