Facultad de INFORMÁTICA

UNIVERSIDAD NACIONAL DE LA PLATA

Servicio de Recolección de Metadatos genérico para documentos

Autor Rodríguez Vuan, Julieta Paz

Directora De Giusti, Marisa Raquel

Asesores Profesionales Villarreal Gonzalo Luján

Lira, Ariel Jorge

Esta obra está bajo una Licencia Creative Commons Atribución-CompartirIgual 4.0 Internacional.

Objetivo

Servicio web que recibe **peticiones** para la **extracción** de información de un artículo científico a través de una **URL**. Al recibir la petición la aplicación **recupera** el documento HTML correspondiente y **analiza** su estructura interna a fin de identificar la mayor cantidad de **metadatos** posibles. Una vez completado el análisis y extracción de metadatos, generará una **representación** interna que utilizará para enviar como **respuesta** a la aplicación solicitante.

Introducción

- Contexto
- Motivación
- Estado del Arte
  - Objeto de estudio
  - Sitios web seleccionados
  - Información a extraer
  - Herramientas de interoperabilidad
  - Técnica utilizada
- Desarrollo
  - Tipo de documentos
  - Pruebas
  - Extensión navegador web Chrome
- Conclusiones
- Trabajo futuro

Artículo Científico

Trabajo de investigación publicado en una revista especializada en un cierto tema. Este tipo de documento tiene como objetivo difundir de manera clara y precisa los resultados de una investigación realizada sobre un área determinada de estudio.

Revista Científica

Publicación, digital o impresa, en la que se difunde el progreso de la ciencia exponiendo los artículos científicos generalmente altamente especializados en un área

Portales de Revistas

Están orientados a la difusión de la investigación y al apoyo de la edición de revistas científicas, tanto en papel como electrónicas

OJS
OPEN JOURNAL SYSTEMS

Sistema de Administración y publicación de revistas y documentos periódicos (Seriados) en Internet.

Ejemplos utilizados

PORTAL DE REVISTAS UNLP

Portal de Revistas de la UNLP (OJS 3)

Portal de Revistas Académicas

Portal de Revistas de la UCR (OJS 2)

Editorial

Tiene como objetivo el de producir, difundir y distribuir obras, sobre las que realiza las tareas de producción y difusión.

Springer Nature

Es una de las editoriales de revistas científicas más influyentes del mundo y pionera en el campo de la investigación abierta. Esta editorial posee sitios mundialmente conocidos en los que publica miles de artículos que ayudan a los investigadores a avanzar en las investigaciones en las que trabajan.

Elsevier España

Edita más de un centenar de revistas, entre las que se encuentran las cabeceras oficiales de más de 70 sociedades científico-médicas; cuenta con un amplio fondo editorial de libros de autores destacados, que conjuntamente con nuevas soluciones online, proporciona a los profesionales de la salud y la investigación científica conocimientos e información de alta calidad y amplia cobertura.

Repositorio institucional

Sitio web destinado a brindar servicios que permitan recopilar, catalogar, difundir y preservar recursos con el objetivo asegurar su accesibilidad y uso a largo plazo

Dspace

Es un software de código abierto pensado para la gestión de repositorios digitales que proporciona distintas herramientas y funcionalidades que permiten satisfacer las diferentes necesidades que requieren las instituciones.

SEDICI

Su misión es albergar, preservar, difundir y dar visibilidad a nivel mundial a toda la producción científica e intelectual de las distintas unidades académicas que la componen

Metadato

Datos que describen otros datos. Datos que guardan información acerca de un elemento

Tipos:

- Administrativos
- Estructurales
- Descriptivos
- Técnicos
- De uso

Descriptivos (Implementación)

Dublin Core

MARC

HTML Meta Tags

Dublin Core

Organización abierta que tiene como objetivo es el desarrollo de estándares de metadatos interoperables

CONTENIDO	PROPIEDAD INTELECTUAL	INSTANCIACIÓN
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Source	Rights	Identifier
Language		
Relation		
Coverage		

Herramientas de interoperabilidad

Interoperabilidad: *“la habilidad de dos o más sistemas o componentes para intercambiar información y utilizar la información intercambiada”*

Z39.50

protocolo para la recuperación de información basado en la estructura cliente/servidor que facilita la interconexión de sistemas informáticos.

SRU & SRW

OAI-PMH

Iniciativa para desarrollar y promover estándares de interoperabilidad que faciliten la difusión de contenidos así como el intercambio de formatos bibliográficos entre distintos repositorios digitales y portales de revistas

SWORD

protocolo liviano diseñado para facilitar el depósito interoperable de recursos principalmente en repositorios, pero potencialmente en cualquier sistema en el que se pretenda recibir contenido de fuentes remotas.

DOI

Es una cadena de caracteres utilizada para identificar la propiedad intelectual en el ambiente digital. Constituye un identificador único y permanente de un recurso y un mecanismo para acceder a ese contenido.

Handle

son identificadores persistentes que surgen para solucionar los problemas que se crean cuando se cambia la ubicación y/o nombre de un objeto digital. El objetivo de un identificador persistente es el de redireccionar a los documentos, aunque estos hayan cambiado su ubicación en la red (cambio de URL).

Web Scraping

Forma de extraer de manera automática datos de un sitio web donde luego de la extracción se tratan esos datos como información

Web Scraper Chrome

Mendeley

Herramienta Desarrollada

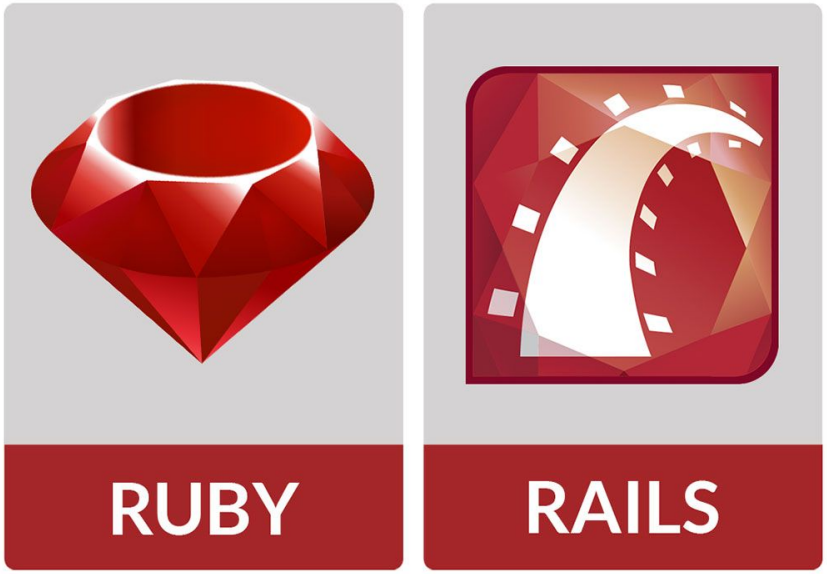
Servicio web capaz de analizar documentos HTML a fin de extraer sus metadatos y retornar estos a la plataforma que lo solicita.

Sencillo

Escalable

Potente

Lenguaje y framework



Nokogiri 鋸

PUMA

PostgreSQL

Documentos

XML

```
<related_url>
  <item>
<url>http://www.publisher.com/url>
    <type>pub</type>
  </item>
  <item>
<url>http://www.author.com</url>
    <type>author</type>
  </item>
</related_url>
```

HTML

```
<head>

  <meta charset="UTF-8">

      <meta    name="description"
content="Tesina de grado">

        <meta     name="keywords"
content="HTML,CSS,XML,JavaScript"
>

      <meta          name="author"
content="Julieta Rodriguez Vuan">

</head>
```

JSON

```
{
          "Equipo":      "SEDICI",
        "ciudad":    "La    Plata",
                "miembros":      [
                                  {
     "nombre": "Julieta Rodriguez
Vuan",
             "puesto":   "Tesista",
                    "trabajo":    [
           "Desarrollo", "Autor",
                                  ]
                                 },
                                  ]
}
```

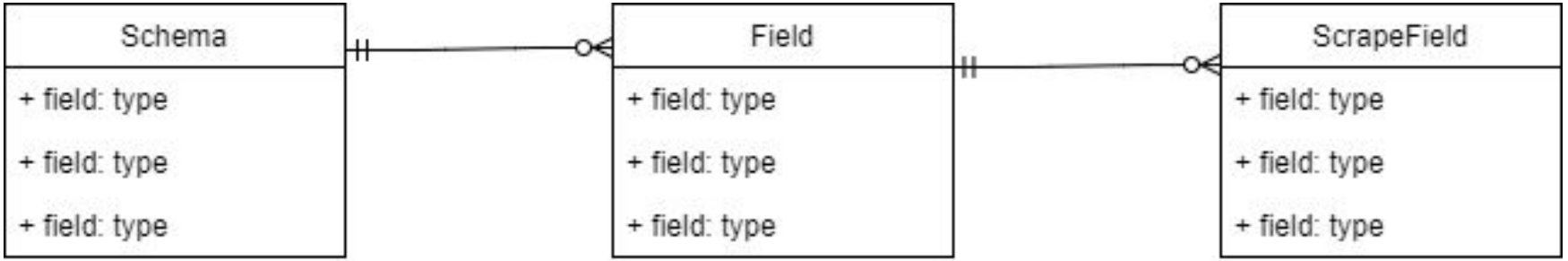
Eprints

DSpace

Patrones de Diseño

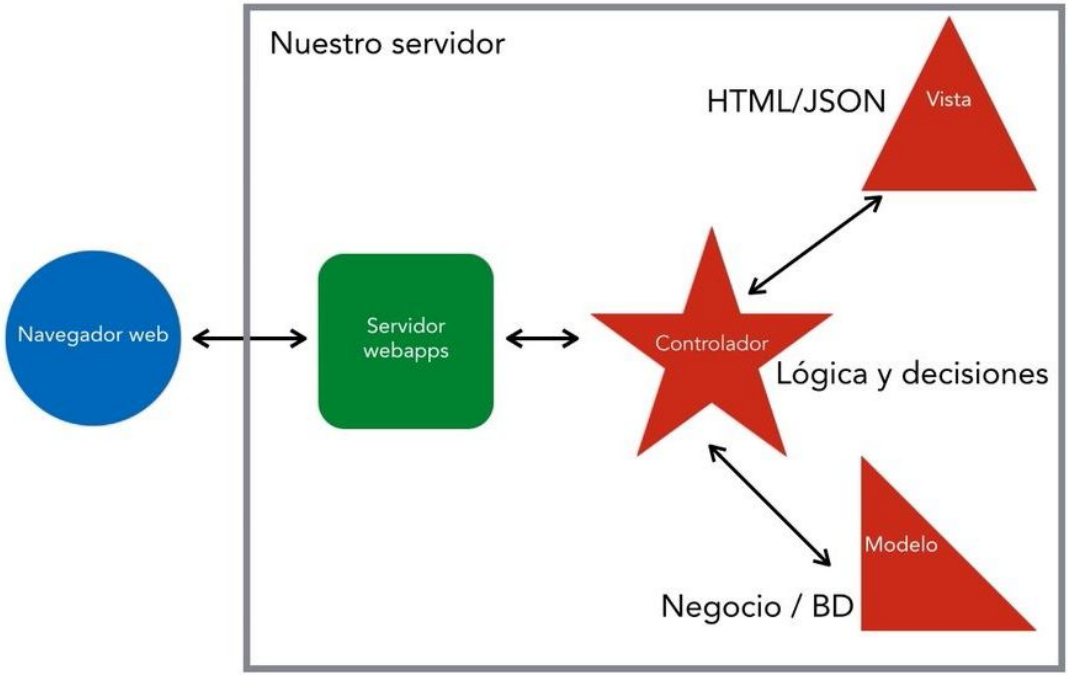
ESTRATEGIA

CADENA DE RESPONSABILIDAD



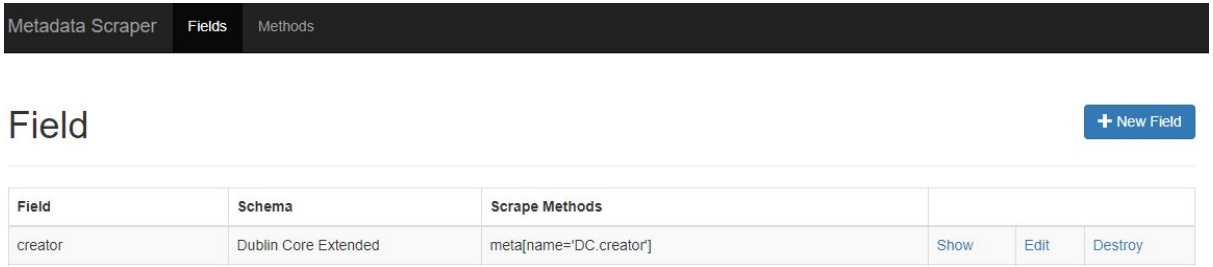
Arquitectura

MVC



Estilo


Metadata Scraper
Fields
Methods
Ingrese su url
Parser de articulos cientificos
Recolectar informacion:
Recolectar


Metadata Scraper
Fields
Methods
Field
+ New Field
Field
Schema
Scrape Methods
creator
Dublin Core Extended
meta[name='DC.creator']
Show
Edit
Destroy

Alcance

Se obtuvo la herramienta que permite el mapeo de metadatos del artículo solicitado al tipo de metadato del sitio que lo solicitó y se desarrolló la extensión del navegador web que personaliza el mapeo al formulario de carga de **SEDICI**

Pruebas

SEDICI

Portal de Revistas UCR

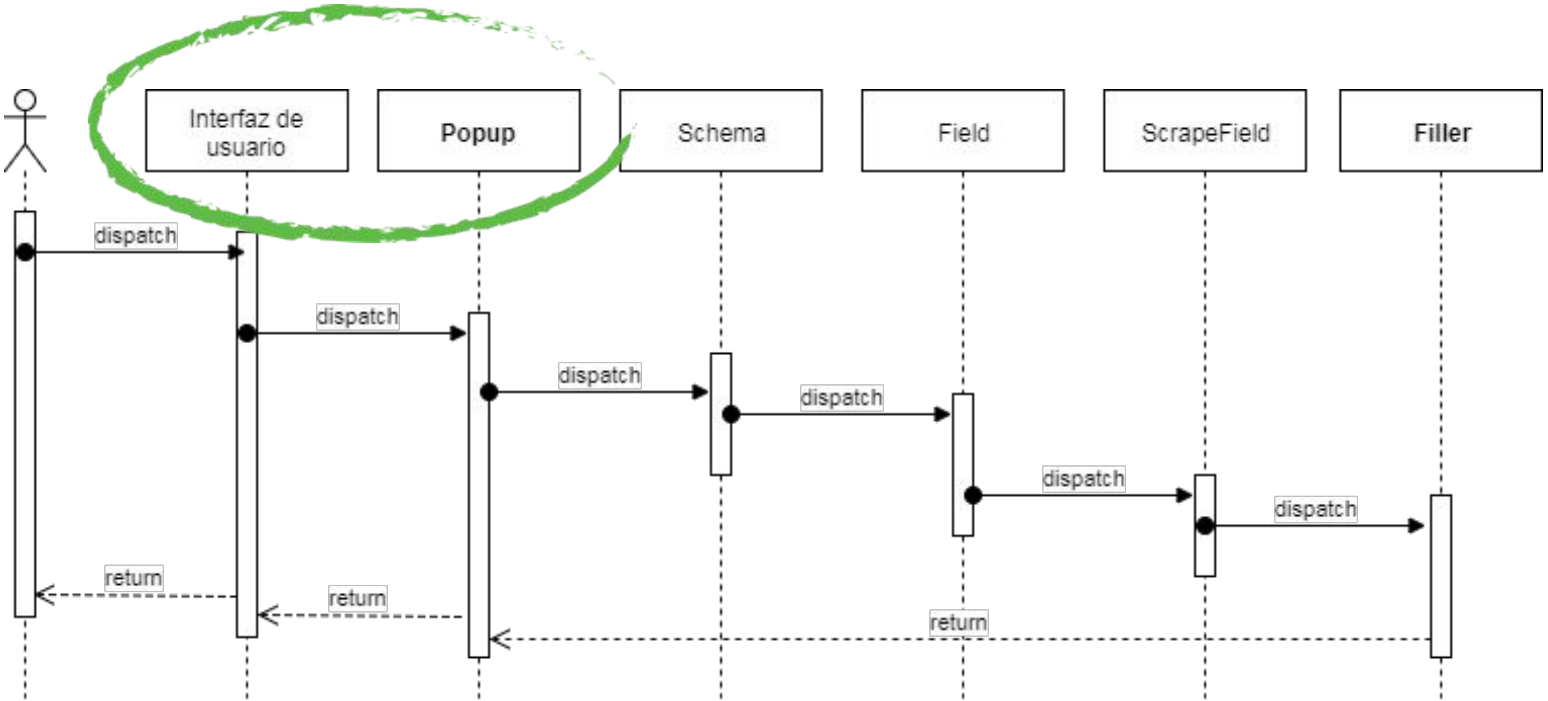
Elsevier España

Extensión navegador Web Chrome

AutoFillMagic_3000 1.2

Carga de formularios automático ;)

Detalles

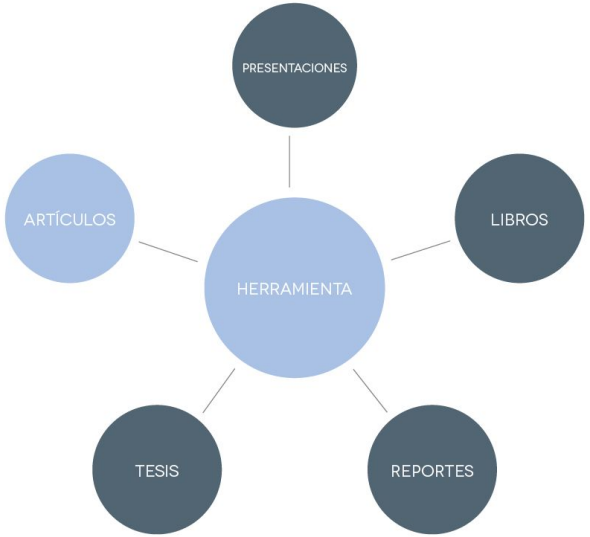


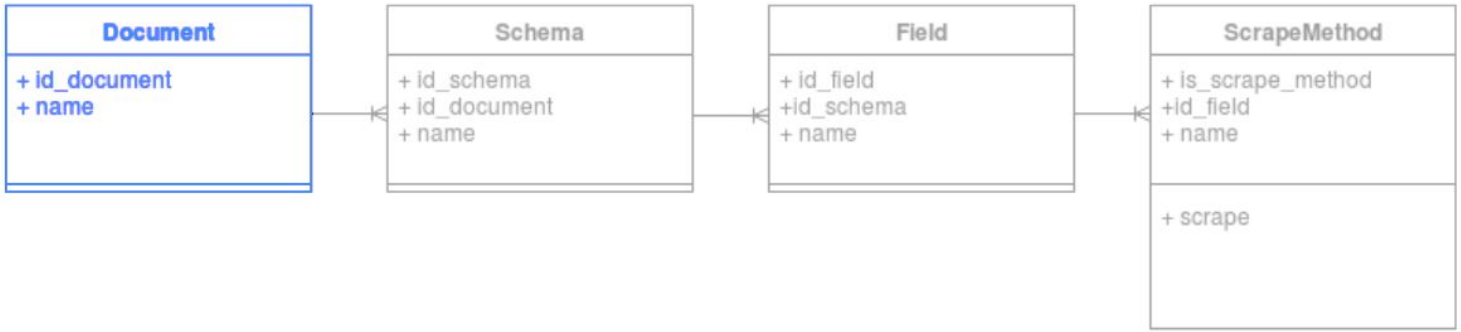
Mejoras

mejora a los metadatos extraídos

resultados en distintos tipos de documentos: CSV, XML

Documentos de distinto tipo





Comparación de metadatos

dc.format.extent	23 p.	es
dc.language	es	es
dc.title	Accesibilidad de los contenidos en un repositorio institucional: análisis, herramientas y usos del formato EPUB	es
dc.type	Articulo	es
sedici.identifier.uri	http://revistas.ucr.ac.cr/index.php/eciencias/article/view/23690	es
sedici.identifier.doi	http://dx.doi.org/10.15517/eci.v6i2.23690	
sedici.identifier.issn	1659-4142	es
sedici.creator.person	De Giusti, Marisa Raquel	es
sedici.creator.person	Lira, Ariel Jorge	es
sedici.creator.person	Rodríguez Vuan, Julieta Paz	es
sedici.creator.person	Villarreal, Gonzalo Luján	es
sedici.subject.materias	Ciencias Informáticas	es
sedici.subject.other	dispositivos móviles	es
sedici.subject.other	accesibilidad	es
sedici.subject.other	nuevas tecnologías	es
sedici.subject.keyword	accesibilidad	es
sedici.subject.keyword	texto a voz	es
sedici.subject.keyword	repositorio institucional	es

creator: "De Giusti, Marisa Raquel;Lira, Ariel Jorge;Rodríguez Vuan, Julieta Paz;Villarreal, Gonzalo Luján"
title: "Accesibilidad de los contenidos en un repositorio institucional: análisis, herramientas y usos del formato EPUB"
title_subtitle: ""
location: ""
extent: "23 p."
abstract: "El objetivo de este trabajo es describir alternativas incorporadas en el formato EPUB3 para promover el acceso a la producción académica y científica de las instituciones por parte de personas con discapacidades visuales. Como punto de partida se toma la figura del repositorio institucional como espacio que alberga y difunde esta producción, y cuyos objetivos incluyen darle mayor visibilidad y maximizar su impacto, manteniéndose así en la misma línea con la propuesta de este estudio. Se analizan los aportes introducidos en el formato EPUB3 con respecto a sus antecesores. En particular, se estudian las extensiones existentes que sirven para optimizar la síntesis de voz a partir de los textos (TTS, text-to-speech), la incorporación de voces adicionales y múltiples voces, y finalmente las herramientas disponibles para visualizar y reproducir documentos EPUB3 con incorporaciones TTS. En este aspecto, se hace énfasis en las aplicaciones accesibles gratuitamente desde dispositivos móviles actuales a fin de asegurar el aprovechamiento de estos aportes por cualquier potencial persona usuaria. Por último, se evalúa la viabilidad de implementar un circuito de generación de documentos EPUB3 accesibles, y se analizan posibles servicios adicionales que el repositorio institucional puede brindar a partir de estas herramientas.;The aim of this work is to describe alternatives introduced in EPUB 3 format to promote access to the academic and scientific institutional production by users with visual disabilities. The figure of the Institutional Repository is taken as starting line, understood as a space which hosts and disseminates this production, and whose objectives include maximizing its impact and fostering its visibility, both in the same line with the proposal of the study. Contributions in EPUB 3 format are analyzed and compared to its predecessors. Extensions for text to speech (TTS) synthesis optimization are studied in depth as well as the ability to add spare and multiple voices, and some of the available software tools to visualize and reproduce TTS-enabled EPUB 3 documents. In this matter, the stress has been put on applications freely available for current mobile devices, in order to ensure that any potential user will be able to take advantage of these contributions. Lastly, the viability of implementing a circuit for accessible EPUB 3 documents generation is discussed, and further services for an institutional repository to offer from these tools are briefly mentioned."
subject: "Ciencias Informáticas;dispositivos móviles;accesibilidad;nuevas tecnologías;accesibilidad;texto a voz;repositorio institucional;EPUB3;accesibility;text-to-speech;institutional repository;EPUB3;Text processing"
document_type: "Articulo;Reporte"
keywords: "Ciencias Informáticas; Text processing; dispositivos móviles; accesibilidad; nuevas tecnologías; Artículo; Reporte; accesibilidad; texto a voz; repositorio institucional; EPUB3; accesibility; text-to-speech; institutional repository; EPUB3"
identifier_uri: "http://hdl.handle.net/10915/53638;http://revistas.ucr.ac.cr/index.php/eciencias/article/view/23690;http://dx.doi.org/10.15517/eci.v6i2.23690;1659-4142"
issn: "1659-4142"
journal_title: "e-Ciencias de la Información"
volume_and_issue: "e-Ciencias de la Información;vol. 6, no. 2"
date_published: "2016-07"

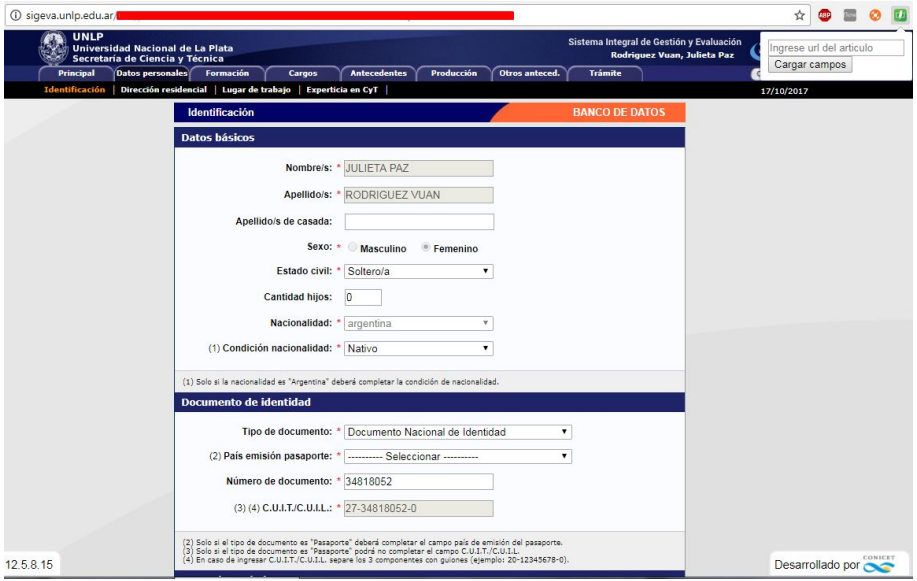
Análisis gramatical de listado de links

Se podría utilizar la tabla de contenidos de una revista como:
revistas.unlp.edu.ar/analecta/issue/current

```
lista = doc.css('div.heat a').map { |link| link['href'] }
```

Automatización de carga

Incrementar la velocidad de carga de datos en los sistemas de información académica



Facultad de INFORMÁTICA

UNIVERSIDAD NACIONAL DE LA PLATA

¡GRACIAS!

¿Preguntas?

Julieta Paz Rodríguez Vuan

Fecha XX de XX de 2017