COAR Confederation of Open Access Repositories

COAR Annual Meeting 2023
Sustainability and Innovation in Scholarly Communications

Mesa

Contenido multilingüe y no inglés en repositorios

Las acciones del repositorio institucional **SEDICI-UNLP** a la luz de la recomendación 1 de COAR sobre multilingüismo

De Giusti Marisa R., Lira Ariel J.

17 de mayo de 2023

PREBI prebi.unlp.edu.ar SEDICI sedici.unlp.edu.ar

UNIVERSIDAD NACIONAL DE LA PLATA

CESGI Centro de Servicios en Gestión de Información

COMISIÓN DE INVESTIGACIONES CIENTÍFICAS

Multilingüismo

- El multilingüismo es una característica fundamental de un panorama de comunicaciones de investigación saludable, inclusivo y diverso.
- El multilingüismo supone un reto especial para el descubrimiento de los resultados de la investigación.
- Los sistemas de búsqueda como **Google Scholar** y otros índices académicos tienden a proporcionar acceso sólo al contenido disponible en la lengua del usuario o en su defecto en inglés.
- Además, el idioma de un recurso académico a menudo no está etiquetado adecuadamente, lo que significa que una gran parte de los recursos que no están en inglés quedan excluidos de los resultados de la búsqueda.
- En agosto de 2022, **COAR** puso en marcha el Grupo de Trabajo de apoyo al multilingüismo.

Casos de uso que guiaron las recomendaciones

Como investigador:

- Quiero encontrar todos los artículos que sean relevantes para mi interés, independientemente del idioma en el que estén publicados.
- Quiero saber si existe una traducción de un artículo o si este documento es una traducción de otro documento.

Como gestor de repositorio:

- Quiero saber cuál es la mejor manera de etiquetar los artículos, tesis o disertaciones que están escritos en más de un idioma para que los lectores sean conscientes de las distintas lenguas.
- Quiero ofrecer metadatos tanto en mi idioma local como en inglés para que el contenido forme parte del registro académico internacional y sea visible para todos.
- Quiero exponer el idioma del artículo en OAI-PMH.

Como agregador:

- Quiero saber cuál es el idioma del documento de texto completo que estoy indexando, para poder ayudar a los usuarios a encontrar el contenido en su idioma preferido.

Escenario de esta presentación

SEDICI
sedici.unlp.edu.ar

Repositorio institucional de la
Universidad Nacional de La Plata, Argentina
Año creación: **2003**

UNIVERSIDAD NACIONAL DE LA PLATA

Total publicaciones: **144000**

Crecimiento prom. mensual (9 años): **1000 items**

Crecimiento en 2022: **17000**

Posición 14 en el *Transparent Ranking* de CSIC en la categoría de repositorios institucionales: **https://repositories.webometrics.info/en/node/32**

RANK	INSTITUTIONAL REPOSITORY	ITEMS
12	Repositório Institucional Universidade Federal de Santa Catarina	
12	UPCommons Universitat Politécnica de Catalunya	
14	Servicio de Difusión de la Creación Intelectual Universidad Nacional de la Plata	
15	Repositório da Produção Universidade de São Paulo	
15	IDUS Depósito de Investigación Universidad de Sevilla	

Recomendación inicial

El 1 de noviembre de 2022, el Grupo de Trabajo COAR publicó su **recomendación inicial para mejorar el descubrimiento de contenidos de los repositorios en una variedad de idiomas, junto con una guía de implementación para la comunidad de repositorios.**

Recomendación 1
Todos los registros en el repositorio deben incluir una etiqueta en el campo de metadatos de idioma que identifique el idioma del recurso y una etiqueta que identifique el idioma de los metadatos (incluso si los recursos están en inglés).

¿Por qué?
Esta es una recomendación muy simple, pero extremadamente poderosa. **Cuando el idioma de los metadatos y el idioma del recurso se atribuyen correctamente, permite que los servicios de descubrimiento e indexación procesen y analicen correctamente el texto.**

Recomendación 1.1
Práctica recomendada para el uso de etiquetas de idioma

Mantener la etiqueta de idioma lo más corta posible. Evitar la región, el guión u otras subetiquetas, excepto cuando se agreguen elementos distintivos e informativos útiles. Por ejemplo, se recomienda usar ja para japonés y no ja-JP

Aplica tanto para el dc.language como para la indicación del metadato con xml:lang o lo que se elija. se debería usar código de idioma lo más corto que sea posible.

En SEDICI se usan 2 letras para el idioma, por ejemplo: http://sedici.unlp.edu.ar/handle/10915/139885?show=full

dc.identifier.uri	http://sedici.unlp.edu.ar/handle/10915/139885
dc.description.abstract	Este trabajo está dedicado a relatar la exper realizada utilizando NDSA, se inicia a partir de la y los estándares utilizados para auditar la con acceso y comprensión de sus contenidos a largo áreas del repositorio, es muy comprensible y bri sencillo que otros estándares como la ISO 1636 repositorio CIC Digital así como una propuesta c o parcialmente cumplidos.
dc.language	es

Recomendación 1.2
Cómo atribuir idioma a un recurso de repositorio en un idioma

Identificar el idioma principal del recurso tanto a nivel de archivo como de ítem.

A. A nivel de archivo: se debe exponer siempre que se da un link o se referencia a un archivo. Ejemplo: con atributos xml:lang
B. A nivel de ítem: se debe exponer siempre que se muestra información de un ítem. Ejemplo: con un tag *<dc:language>**fr**</dc:language>*

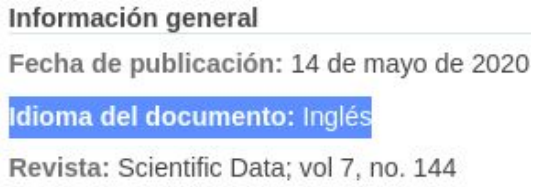
Recomendación 1.2 en SEDICI

A) A nivel de ítem

- Cada recurso se expone como un ítem en un único idioma, es decir, una tesis cuando está en d**os idiomas** se guarda en **dos ítems distintos.**
- Para cada ítem se mantiene un metadato *dc:language* en la base de datos

Ej. en **web, para el usuario**:

http://sedici.unlp.edu.ar/handle/10915/139885



Ej en **web, para Google Scholar**:

<meta content="es" name="citation_language" />

Ej en OAI

<dc:language>**es**</dc:language>
http://sedici.unlp.edu.ar/oai/request?verb=ListRecords&metadataPrefix=oai_dc

Recomendación 1.2 en SEDICI

B) A nivel de archivo

A nivel de archivo no lo tenemos pero podríamos exponerlo, ya que en general tenemos un solo idioma por ítem y esto hace fácil inferir el idioma de cada archivo.

- En interfaz web, para bots y agentes de usuario (navegadores, readers, etc): Se podría agregar un atributo *xml:lang* a la etiqueta <A> que apunta a cada archivo/bitstream.

   Descargar archivo

- En OAI-PMH: se puede agregar el atributo *xml:lang* en los enlaces a los archivos. Ej en oai bajo directrices OpenAire4 sería:

  <oaire:file … objectType="fulltext" xml:lang="es">
  http://link-to-the-fulltext.org/bitstream/handle/10915/152880…
  </oaire:file>

Recomendación 1.3

Cómo atribuir un lenguaje a un recurso que tiene más de un lenguaje

Use el atributo xml:lang y la etiqueta dc:language para indicar los idiomas de los recursos.

Vea los siguientes ejemplos:

<oaire:file accessRightsURI="https:/example.com/english.pdf" mimeType="application/pdf" objectType="fulltext" xml:lang="en">http://link-to-the-fulltext.org</oaire:file>

<oaire:file accessRightsURI="https:/example.com/french.pdf" mimeType="application/pdf" objectType="fulltext" xml:lang="fr">http://link-to-the-fulltext.org</oaire:file>

AND

*<dc:language>**en**</dc:language>*

*<dc:language>**fr**</dc:language>*

Recomendación 1.3 en SEDICI

En SEDICI no suele pasar este caso, ya que cada ítem tiene un idioma y los archivos son del mismo idioma (como en 1.2). Se podría agregar la posibilidad de cargar múltiples idiomas para un ítem dado, sin problemas.
Sin embargo, no podríamos registrar correctamente el idioma de archivos (bitstreams) con diferentes idiomas ya que no se guarda el idioma del archivo en la base de datos.

En **DSpace** sólo se guarda el idioma por ítem, pero no por bitstream. Entonces:

a) habría que modificar el código de DSpace para permitir el idioma de un archivo/bitstream y guardar el valor en un metadato del mismo, o
b) se podría indicar múltiples idiomas en los metadatos del ítem, pero sin indicar a qué archivo corresponde cada uno.

Recomendación 1.4

Cómo atribuir el idioma cuando hay más de un idioma en el campos de metadatos

Utilice el atributo xml:lang para indicar el idioma del campo de metadatos. Si esto está vacío, los servicios de descubrimiento asumirán que los metadatos están en el mismo idioma que el recurso

Vea los siguientes ejemplos:

<datacite:titles>
<datacite:title xml:lang="en">**Open Access**</datacite:title>
<datacite:title xml:lang="pl">**Otwarty Dostęp**</datacite:title>
</datacite:titles>
<dc:title xml:lang="en">**Open Access**</dc:title>
<dc:title xml:lang="fr">**Libre Accès**</dc:title>

Recomendación 1.4 en SEDICI

dc.description.abstract	Este trabajo está dedicado a relata realizada utilizando NDSA, se inicia a y los estándares utilizados para aud acceso y comprensión de sus conteni áreas del repositorio, es muy compre sencillo que otros estándares como la repositorio CIC Digital así como una p o parcialmente cumplidos.	es
dc.language	es	es
dc.subject	repositorio	es
dc.subject	confiabilidad	es
dc.subject	autoealuación	es
dc.subject	NDSA	es
dc.title	Evaluación de CIC Digital a través de	es
dc.type	Objeto de conferencia	es

En SEDICI y en cualquier DSpace se puede guardar el idioma de cada metadato.

El idioma del metadato se expone solo para google scholar, pero podría indicarse tanto en el OAI como en la interfaz de usuario usando el atributo **xml:lang.**

Ej en OAI

<dc:subject xml:lang="es">**ingeniería vial**
</dc:subject>

En web

<span class="metadata-value"
*xml:lang="es">**ingeniería vial***

Sugerencia: relación entre traducciones

Artículo
The TRUST Principles for digital repositories

dc.identifier.uri	http://sedici.unlp.edu.ar/handle/10915/96052
dc.language	en
dc.subject	digital data
dc.title	The TRUST Principles for digital repositories
dc.type	Articulo
sedici.identifier.other	DOI:10.1038/s41597-020-0486-7
sedici.identifier.issn	2052-4463
sedici.relation.isRelatedWith	http://sedici.unlp.edu.ar/handle/10915/97465

Artículo
Los principios TRUST en los repositorios digitales

dc.identifier.uri	http://sedici.unlp.edu.ar/handle/10915/97465
dc.language	es
dc.title	Los principios TRUST en los repositorios digitale
dc.type	Articulo
sedici.identifier.issn	2052-4463
sedici.description.note	Este artículo es una traducción de: Lin, D., C Edmunds, R., Giaretta, D., De Giusti, M., L' Khodiyar, V., Martone, M. E., Mokrane, M., N Sokolova, D. V., Stockhause, M., & Westbrook, digital repositories. Scientific Data, 7(1), 144. 0486-7
sedici.relation.isRelatedWith	http://sedici.unlp.edu.ar/handle/10915/96052

- **En SEDICI** usamos una relación libre pero no es lo suficientemente expresiva:
  - Versión en español http://sedici.unlp.edu.ar/handle/10915/97465
  - Versión en inglés http://sedici.unlp.edu.ar/handle/10915/96052

Sugerencia: relación entre traducciones

Deberíamos guardar metadatos de relación entre traducciones de una obra.

- Las relaciones serían 2:
  - **A** es traducción de **B**
  - **B** es traducida por **A**

- En **Dublin core** no hay nada específico (https://www.dublincore.org/specifications/dublin-core/relation-element/)
  - *IsBasedOn / IsBasisFor:* Creative relations are those in which one resource is a performance, production, derivation, translation, adaptation or interpretation of another resource. The corresponding values of Relation.Type are:
- En **datacite** tampoco https://support.datacite.org/docs/connecting-to-works

- **Opciones simples para exponer idioma en relaciones**:
  - <a xml:lang="es" href="http://sedici.unlp.edu.ar/handle/10915/96052">Traducción al español **(en HTML para interfaz web)**
  - <link rel=”alternate” lang="fr" href="http://…/mydoc-fr.html"> **(en HTML para interfaz web)**
  - <dc:relation xml:lang=”fr”>http://…/mydoc-fr.html</dc:language> **(en XML para OAI)**

Problemas pendientes

Como investigador:

- **Problema 1:** quiero descubrir/encontrar material de interés que no está ni en mi idioma nativo ni en inglés
- **Problema 2:** cómo encontrar o generar una traducción de un trabajo que no está en mi idioma.
- **Problema 3:** cómo publicar en mi idioma nativo (no inglés) y a pesar de eso que los mismos impacten en el público global

- Todos son difíciles de resolver y aceptan múltiples enfoques que seguramente se irán desarrollando y perfeccionando en el tiempo
- Lo importante **hoy** es dar soporte desde los metadatos para que otros implementen los servicios en el futuro

¡Muchas gracias!

marisa.degiusti@sedici.unlp.edu.ar
alira@sedici.unlp.edu.ar

CIC-DIGITAL https://digital.cic.gba.gob.ar/
SEDICI http://sedici.unlp.edu.ar/